THE UNIVERSITY OF ALBERTA

RELEASE FORM

NAME OF AUTHOR ....RANGASWAMI GEETHA..........

TITLE OF THESIS ....PARAMETRIC AND NONPARAMETRIC DISCRIMINANT..........

..........ANALYSIS..........

DEGREE FOR WHICH THESIS WAS PRESENTED ..M.Sc...........

YEAR THIS DEGREE GRANTED ...1975..........

THE UNIVERSITY OF ALBERTA

PARAMETRIC AND NONPARAMETRIC DISCRIMINANT ANALYSIS

BY

RANGASWAMI GEETHA

A THESIS

SUBMITTED TO THE FACULTY OF GRADUATE STUDIES AND RESEARCH

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE

OF MASTER OF SCIENCE

IN

MATHEMATICAL STATISTICS

DEPARTMENT OF MATHEMATICS

EDMONTON, ALBERTA

FALL, 1975

THE UNIVERSITY OF ALBERTA

FACULTY OF GRADUATE STUDIES AND RESEARCH

The undersigned certify that they have read, and recommend to the Faculty of Graduate Studies and Research, for acceptance, a thesis entitled "PARAMETRIC AND NONPARAMETRIC DISCRIMINANT ANALYSIS" submitted by RANGASWAMI GEETHA in partial fulfilment of the requirements for the degree of Master of SCIENCE.

ABSTRACT

For statistical classification or discrimination among two or more classes, the present study discusses various available parametric and nonparametric classification procedures and their associated probability of correct classification, all procedures being derived from a single statistical perspective namely by maximizing rather obvious estimators of probability of correct classification. The study includes some general remarks on these classification procedures.

ACKNOWLEDGEMENT

I wish to thank Professor K.L. Mehra for his encouragement and valuable suggestions, during the preparation of this thesis. I would also like to thank the University of Alberta for financial assistance and Miss Olwyn Buckland for excellent typing of this thesis.

TABLE OF CONTENTS

CHAPTER I

Introduction

The object of this study is to look into and present certain aspects of the classification problem, including various classification procedures discussed in the literature. The problem of Discrimination, or also known as the Identification problem, concerns itself with correctly allocating an individual into one of a specified number k of populations. The Classification problem, on the other hand, is concerned with classifying a sample of individuals into groups, which are to be distinct in some sense. These two problems basically are virtually the same in nature. We shall accordingly be using in the sequel the terms discrimination / allocation / identification / assignment as synonyms in reference to the same classification problem.

In principle, the classification problem is one of the simplest in statistics; in practice however, it has a large number of snags, largely because the assumed theoretical model does not always reflect the practical situation sufficiently closely. The problem was considered to be of practical importance as early as 1935. "Classification" has application in medical diagnosis and treatment, in drug interaction studies, neurobiological signal processing, sonar detection etc. Clinical data, such as electro-cardiograms and electro-encephalograms, can also be analysed and classified using classification techniques. Besides medical problems, other familiar instances where such a problem arises are:

(i) When an anthropologist faces the problem of sexing the skull or jawbone;

(ii) When a taxonomist is assigned the problem of classifying an organism into species or subspecies;

(iii) Authorship of a disputed article; etc...

Among the well-known classification procedures developed, are Fisher's linear discriminant function introduced by Fisher [1936] and Anderson's classification statistic introduced by Anderson [1951]. It was Welch [1939], who gave the first mathematical formulation, on the basis of the foundations laid by Neyman and Pearson in the theory of testing of hypotheses. Subsequent authors made many refinements giving different classification statistics. Rao [1969], in his paper, considers the extended formulation of the classification problem that recognises the possibility of an individual belonging to an unspecified population, as for example, when a biologist discovers a member of a species. In this connection, Srivastava [1973] proposed the "step-down" procedure for classification into one of two multivariate normal populations. Relatively very little has been done in the area of multiple group discrimination. Only recently, Lachenbruch [1973] has proposed two methods for classification into one of several populations and has studied their relative performance. The estimation of probabilities of misclassification has been studied in detail by Dunn and Varady [1966], Hills [1966]. In this connection, among others the papers by Glick [1972] and Lachenbruch and Mickey [1968] should be mentioned.

In Chapter II, we give a detailed account of all available major parametric classification procedures. Section 2.2 deals mainly with rules of classification into known distributions, including the well-known Fisher's linear discriminant function rule and Mahalanobis' generalized squared distance rule. Sample-based classification rules are dealt with in Section 2.3. These arise when the distributions are not specified completely and information on them is to be obtained from the samples. The chapter includes expressions for the optimal probability of correct classification. A review of the literature dealing with these classification rules and the associated probabilities of misclassification is also given.

Chapter III deals with the non-parametric classification problem. The required estimation of probability density functions in such problems has been discussed in detail under section 3.2. The problem of density estimation has received attention only recently in the literature. Fixed window density estimates were suggested by Parzen [1962] and Cacoullos [1966]. The section includes Loftsgaarden and Quesenberry's [1965] fixed view density estimation method as well. In Section 3.3, different non-parametric classification procedures available in the literature are discussed. These rules include the nearest neighbor rule suggested by Fix and Hodges [1951], minimum distance classification rule as suggested by Das Gupta [1964], the best-count rule proposed by Glick [1969] and a few others.

Chapter IV deals mainly with the mathematical proofs of various assertions, made in Chapter II and Chapter III on parametric and

non-parametric classification procedures, and the associated probability of correct classification.

Finally, in Chapter V, we make some general remarks on Classification theory which may be of importance in applications and further research work.

(For computational examples, see Appendix I.)

CHAPTER II

Parametric Classification

In this chapter, we introduce some major parametric rules of classification into known distributions and sample-based classification rules. All these rules assume the existence of underlying densities, with parameters known or unknown. In the case of unknown parameters, simple estimates of parameters prove helpful for the construction of classification procedures. We also study in brief the probabilities of correct classification discussed in the literature.

§2.1 Main Formulations of the Problem.

(i) Let $\pi_1,\pi_2,\ldots,\pi_k$ be k distinct populations (groups/categories/classes). Given a random sample from an unknown population π_o, but known to be one of $\pi_1,\pi_2,\ldots,\pi_k$, the problem of classification demands a decision, as to which one of the latter k populations is π_o, that is optimum in some sense. Since a decision rule is a function from the sample space, $\mathcal{X}$, to the set of decisions, $\pi_1,\pi_2,\ldots,\pi_k$, it will be based upon the observation vector $\underset{\sim}{X}$, and the available information about the distributions π_i $(i = 1,2,\ldots,k)$. If the information is unspecified or inadequate, supplementary information can be obtained through random samples from each of the k populations; such samples being termed "training" samples.

(ii) Suppose there is a population Γ, consisting of k mutually exclusive subpopulations $\pi_1,\pi_2,\ldots,\pi_k$ mixed in respective proportions

(a priori probabilities) $q_1,q_2,\ldots,q_k$ $(q_i \geq 0 , 1 \leq i \leq k , \sum_{i=1}^{k} q_i = 1)$, known or unknown. An individual selected at random from Γ may be regarded as a random vector $<I,X>$, where I denotes the individual's group, and X is the p - dimensional vector of measurements. For the units to be classified, I is unobservable, but X can be observed. The problem of classification amounts to making an inference on the value of I from the knowledge of X . The distribution of I is over the set $\{1,2,\ldots,k\}$. The problem will be termed as the "known mixture" or "unknown mixture" problem according as the distribution of I is known or unknown.

In constructing a classification procedure, it is desired to minimize the expected losses or the probabilities of misclassifying an individual. A procedure which achieves this minimum is called the "best" or "optimal" procedure.

Remark 2.1: In the preceding formulation, one may consider, more generally, I as a continuous or discrete variable with a physical meaning, and the population π_i corresponds to $I \epsilon S_i$ where $S_1,S_2,\ldots,S_k$ is a partition of the I - space. Marshall and Olkin [1968] include the decision of observing I along with making k decisions in their formulation.

§2.2 Classification into Known Distributions.

2.2.1 Bayes Procedure.

Consider the formulation (ii) of section 2.1 with q_i's $(1 \leq i \leq k)$ known. On the basis of the observed $X = x$, a decision, optimum in the sense described in section 2.1, has to be reached about the membership of the individual in one of k specified populations.

The probabilistic structure may be specified by

$$P\{I=i\} = q_i \quad , \quad 1 \leq i \leq k$$

$$P[X<x|I=i] = F_i(x) \quad , \quad 1 \leq i \leq k \ , \ x \in \mathcal{X} \ .$$

A nonrandomized decision rule D consists of the partition of the sample space $\mathcal{X}$ into k mutually exclusive regions $D_1, D_2, \ldots, D_k$, with a rule which assigns an individual, with measurement vector $\underset{\sim}{X} = x$, into the ith population if and only if the observed $x \in D_i$, $i = 1,2,\ldots,k$. Let D^* denote the collection of all classification rules.

Since the number of decisions (classifications) is finite, attention may be restricted to nonrandomized decision rules. It is well known that, in a finite decision problem the optimal solutions for the randomized and nonrandomized rules are essentially the same. (For the definition of the randomized rules and the proof of this assertion, see Rao [1973] section 7d.3.)

Let $f_1, f_2, \ldots, f_k$ denote the probability densities of $F_1, F_2, \ldots, F_k$ respectively, with respect to a σ - finite measure μ . Suppose further, that a loss, $C_{ij} (>0)$, is incurred in assigning an individual from the ith population to the jth population. A loss function which assigns 0 loss to correct classification, and unit loss to any misclassification, is called a simple loss function, i.e., for a simple loss function

$$C_{ij} = \begin{cases} 0 & \text{if} \quad i = j \\ 1 & \text{if} \quad i \neq j \ . \end{cases} \tag{2.2.1}$$

For a nonrandomized rule, the expected loss in applying a given rule D , when in fact the individuals belong to the ith population is

$$L_i = \sum_{j=1}^{k} \int_{D_j} C_{ij} \, f_i(x) \, d\mu(x) \quad , \quad i = 1,2,\ldots,k \ .$$

Knowing the prior probabilities q_i , $1 \leq i \leq k$, the expected loss of incorrectly classifying an individual from the mixed population Γ , associated with D , is

$$\begin{aligned} \rho(D) &= \sum_{i=1}^{k} q_i \, L_i \\ &= \sum_{j=1}^{k} \{-\int_{D_j} g_j(x) \, d\mu(x)\} \quad , \end{aligned} \tag{2.2.2}$$

where

(2.2.3) $$g_j(x) = - \sum_{i=1}^{k} q_i C_{ij} f_i(x)$$

is the so called jth discriminant score of an individual, $1 \leq j \leq k$.

Define by $\gamma(x)$ the maximum of the discriminant scores:

(2.2.4) $$\gamma(x) = \max_{1 \leq j \leq k} g_j(x) \quad .$$

The Bayes Rule D^* corresponding to a given a priori distribution $\{q_1, q_2, \ldots, q_k\}$ always exists and consists of assigning an individual to that population for which his discriminant score, defined by (2.2.3), is the highest. (For a proof, see Rao [1973] p. 493 result (i); or Anderson [1958] section 6.6.) An optimal partition corresponding to this Bayes rule D^* is expressible as $D^* = \{D_1^*, D_2^*, \ldots, D_k^*\}$ where

(2.2.5) $$D_j^* = \{x \in \mathcal{X} : g_j(x) = \gamma(x)\} \quad , \quad 1 \leq j \leq k \quad .$$

Ties may be resolved arbitrarily, e.g., specify a unique partition by taking $x \in D_j^*$ if and only if j is the smallest integer for which the maximum is attained.

As a particular case, consider the problem of classifying an individual into one of two specified populations: i.e., $k = 2$. By the preceding arguments, the classification problem amounts to determining two regions, D_1^* and D_2^* , which minimize the expected loss (2.2.2). The optimal rule is

$$(2.2.6)\quad \begin{cases} D_1^* = \{x \in \mathcal{X} : \dfrac{f_1(x)}{f_2(x)} > C\} \\ \\ D_2^* = \{x \in \mathcal{X} : \dfrac{f_1(x)}{f_2(x)} < C\} \end{cases} ,$$

where $C = \dfrac{C_{21}q_2}{C_{12}q_1}$ depends on the relative losses of misclassification and the prior probabilities. The case when $f_1(x) = C\, f_2(x)$ can be resolved in some arbitrary manner, such as flipping a coin and deciding that an individual comes from π_1 or π_2 according as the coin shows a head or tail.

Remark 2.2. (i) Let $\rho^* = \inf_{D \in \mathcal{D}^*} \rho(D)$.

Then the optimal Bayes Rule D^* is the one which minimizes $\rho(D)$; i.e. the optimal Bayes rule D^* satisfies

$$\rho(D^*) = \rho^* .$$

We call ρ^* the Bayes risk.

(ii) In many practical problems, it is difficult to assess the losses due to wrong classification. In such cases, simple loss structure is assumed and L_i (2.2.2) represents the expected proportion of wrong identifications for individuals of the ith population. So the criterion of minimizing the probabilities of misclassification may serve the purpose, and $g_j(x)$, the jth discriminant score defined by (2.2.3), reduces to

$$g_j(x) = - \sum_{\substack{i=1 \\ i \neq j}}^{k} q_i \; f_i(x)$$

$$= - \sum_{i=1}^{k} q_i \; f_i(x) + q_j \; f_j(x)$$

$$= \text{const} + q_j \; f_j(x) \quad ,$$

i.e. for this purpose, $g_j(x)$ may simply be defined as $q_j \; f_j(x)$.

2.2.2 Minimax Rule.

In the preceding section, the formulation (ii) was considered with the q_i's , $1 \leq i \leq k$, known. It was seen that the optimal Bayes rule D^* depends upon the prior probabilities $q_1, q_2, \ldots, q_k$. In most instances of classification problem, prior probabilities $q_1, q_2, \ldots, q_k$ are not known to the statistician. Rao [1969] has suggested the maximum likelihood method for estimating these q_i's , $1 \leq i \leq k$. Such a problem of unknown prior probabilities arises, for example, in the case of differential diagnosis of diseases, where the diseases may exhibit seasonal variations. It is not possible, in such cases, to implement an optimal rule that minimizes the expected loss. Instead, one minimizes the maximum risk. This criterion is the so-called Minimax Criterion. The determination of such a rule, even if it exists, may be difficult. But there exist situations where a decision rule may be identified as a minimax rule. It has been proved that minimax procedures are Bayes solutions with respect to a least favourable 'a priori' distribution, and the minimax risk equals the so called

maximum Bayes risk. More generally, if there exists no such prior distributuion but only a sequence for which the Bayes risk tends to the maximum, then the minimax procedures are limits of the associated sequence of Bayes solutions (see Lehmann [1959] p. 17, or Rao [1973] p. 496).

2.2.3 Linear Discriminant Function Rule:

The linear discriminant function rule (LDF rule), for classifying an individual into one of two multivariate normal populations with the same covariance matrix, was first introduced by Sir Ronald Fisher in 1936. Fisher's idea was the basis for most of the research in multivariate statistical classification theory. The method of finding discriminant functions in arriving at test criteria for classification procedures has been found extremely useful in multivariate analysis.

Suppose the populations have multivariate normal distributions with the same covariance matrix Σ, but different mean vectors. The ith density $(i=1,2)$ is given by

$$f_i(x) = (2\pi)^{-p/2} |\Sigma|^{-1/2} \exp\{-\tfrac{1}{2}(X-\mu^{(i)})' \Sigma^{-1}(X-\mu^{(i)})\} \quad ,$$

where $\mu^{(i)}$ $(i=1,2)$ denotes the mean vector of the two populations. The ratio of the densities is

$$\frac{f_1(x)}{f_2(x)} = \frac{\exp\{-\frac{1}{2}(X-\mu^{(1)})' \Sigma^{-1} (X-\mu^{(1)})\}}{\exp\{-\frac{1}{2}(X-\mu^{(2)})' \Sigma^{-1} (X-\mu^{(2)})\}}$$

(2.2.7)

$$= \exp\{-\tfrac{1}{2}[(X-\mu^{(1)})' \Sigma^{-1}(X-\mu^{(1)})-(X-\mu^{(2)})' \Sigma^{-1}(X-\mu^{(2)})]\} \quad .$$

Invoking the Bayes classification procedure for the case $k = 2$ (see (2.2.6)), the region of classification into π_1, D_1^*, is the set of X's for which the right hand side of (2.2.7) is greater than C. The monotonicity of the logarithmic function yields (by rearrangement),

$$(2.2.8) \qquad D_1^* = \{X \in \mathcal{X} : U \equiv X' \Sigma^{-1}(\mu^{(1)}-\mu^{(2)}) - \tfrac{1}{2}(\mu^{(1)}+\mu^{(2)})' \Sigma^{-1}(\mu^{(1)}-\mu^{(2)}) > \log C\} \ .$$

The first term, $X' \Sigma^{-1}(\mu^{(1)}-\mu^{(2)})$, is the well-known Fisher's linear discriminant function, a function linear in the components of the observation vector X.

In the special case in which the two populations are equally likely, and the losses due to misclassification are equal, $C = 1$ (see (2.2.6)), and $\log C = 0$. Then the region of classification into π_1, D_1^*, is

$$D_1^* = \{X \in \mathcal{X} : X' \Sigma^{-1}(\mu^{(1)}-\mu^{(2)}) > \tfrac{1}{2} (\mu^{(1)}+\mu^{(2)})' \Sigma^{-1}(\mu^{(1)}-\mu^{(2)})\} \ .$$

If the a priori probabilities are unknown, we select $\log C = k$, say, by making the expected losses due to the wrong classifications equal. This demands the knowledge of the distribution of U. Anderson [1958] and subsequently many authors studied the distribution of U. It is well-known that U is distributed as $N(\frac{\alpha}{2}, \alpha)$ when X is distributed according to $N(\mu^{(1)},\Sigma)$. When X is distributed according to $N(\mu^{(2)},\Sigma)$, U is distributed as $N(-\frac{\alpha}{2}, \alpha)$, where

$$\alpha = (\mu^{(1)}-\mu^{(2)})' \Sigma^{-1} (\mu^{(1)}-\mu^{(2)}) \quad .$$

The probabilities of misclassification are (see Anderson [1958])

$$P(2|1) = \int_{-\infty}^{(k-\alpha/2)/\sqrt{\alpha}} \frac{1}{\sqrt{2\pi}} e^{-y^2/2} \, dy$$

and

$$P(1|2) = \int_{(k+\alpha/2)/\sqrt{\alpha}}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-y^2/2} \, dy \quad .$$

Thus, for the minimax solution, we choose k so that

$$C_{21} \int_{(k+\alpha/2)/\sqrt{\alpha}}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-y^2/2} \, dy = C_{12} \int_{-\infty}^{(k-\alpha/2)/\sqrt{\alpha}} \frac{1}{\sqrt{2\pi}} e^{-y^2/2} \, dy \quad .$$

A special representation of the probability of correct classification by the optimal LDF rule is given in section 2.2.6. Marshall and Olkin [1968] derived Bayes rule for the normal populations in their special set-up, pointed out earlier. Further, Anderson and Bahadur [1962] considered the problem when the two multivariate normal populations have unequal covariance matrices. The likelihood-ratio method can still be used but it does not lead to a linear discriminant function. The discriminant score for the ith population is, $(i=1,2,\ldots,k)$

$$g_i(x) = -\frac{1}{2} \log |\Sigma_i| - \frac{1}{2} (X-\mu^{(i)})' \Sigma_i^{-1}(X-\mu^{(i)}) + \log q_i$$

which may be called a quadratic discriminant score. The decision rule amounts to assigning an individual to that population for which his

quadratic discriminant score is the highest. Anderson and Bahadur [1962] showed that no linear discriminant function can be an optimal rule. They derived the minimax rule and characterized the minimal complete class. After restricting to the class of rules based on linear functions of X, they also established that among all the linear functions, Fisher's LDF minimizes the probabilities of misclassification. Not much has been studied on nonlinear discriminants subsequent to their paper.

Remark 2.3. (i) The choice of discriminant function in the preceding discussions is not unique. We can always multiply a discriminant function by a positive constant, or bias it by an additive constant without influencing the decision. Consequently, all the decision rules so obtained are equivalent.

(ii) The extension of the above classification problem to classification into one of several multivariate normal populations is discussed in detail in Anderson [1958]. The underlying idea in his approach is the same; namely, an ordered partition of the sample space $\mathcal{X}$ such that the expected loss is a minimum. For a detailed discussion of the topic, one is referred to Anderson [1958, pp. 147].

2.2.4 Minimum Distance Rule.

Consider the formulation (i). So far, in all the above classification procedures it was assumed that the individual to be classified belongs to one of the several specified populations. This assumption is realistic in many taxonomic problems such as sexing of skeletal remains,

where the possibilities of identification is limited to two. However, when the external evidence is slight, the classification is subject not only to error due to misclassification, but also due to the possibly erroneous assumption that it belongs to one of the specified populations. In order to have a better justification of the classification, the best procedure would be to first test whether or not it belongs to one of the given populations. Unfortunately, no such test criterion is available. Alternatively, we find which of the k populations is "nearest" or "closest", measured in terms of some distance function, to the individual to be classified.

An example in which the usual classification approach is not pertinent is the following:

Suppose a relatively new language is to be compared with two or more older languages: The purpose is to find which of these languages is most similar to the former. If a measure of dissimilarity in terms of a distance function between two languages is available, then the question of the nearest to the new one is quite appropriate.

This leads to the question of what measure of distance should be used. For the case of multivariate normal populations, Mahalanobis [1936] proposed the generalized squared distance as a measure of divergence between the populations. The divergence is given by

$$\Delta_p^2 = \sum_{i=1}^{p} \sum_{j=1}^{p} \alpha^{ij} \delta_i \delta_j \quad , \tag{2.2.9}$$

where δ_i denotes the difference in true mean values for the ith

variable, (α^{ij}) denotes the elements of the inverse matrix of the common or pooled covariance matrix, and p in the subscript denotes the number of variables used.

Translating (2.2.9) into matrix notation, we have

$$\Delta_p^{\,2} = (\mu^{(i)}-\mu^{(j)})' \Sigma^{-1} (\mu^{(i)}-\mu^{(j)}) \quad . \tag{2.2.10}$$

Mahalonobis' method is one of the earliest suggested distance methods, having numerous applications in anthropometric studies. This method has become a powerful tool in statistical and biometric research. But, unfortunately, the formula (2.2.9) (or (2.2.10)) is not of much use in practice, since the computation of the inverse matrix and quadratic form in the differences of the mean values becomes extremely laborious when the number of characters exceeds 4 or 5.

As the name suggests, the minimum distance rule classifies an observation into that population which is at a minimum distance. In case of ties, one can make a randomized decision. Consequently, the so called minimum distance rule classifies an observation X_o into π_1 or π_2 (two multivariate normal populations with common covariance matrix Σ) according as

$$(X_o-\mu^{(1)})' \Sigma^{-1}(X_o-\mu^{(1)}) \lessgtr (X_o-\mu^{(2)})' \Sigma^{-1}(X_o-\mu^{(2)}) \quad . \tag{2.2.11}$$

2.2.5 Probability of Correct Classification.

In the classification procedures discussed in the preceding sections, the fundamental criterion for obtaining the optimal rule was to minimize the expected loss or the probabilities of misclassification. Given a rule D , the probability that it will correctly classify an individual chosen randomly from the ith population, is

$$\int_{D_i} d\, F_i(x) = \int_{D_i} f_i(x)\, d\mu(x) \quad , \quad 1 \leq i \leq k \quad .$$

Consequently, the probability that a given rule D will correctly classify an individual selected at random from the mixed population Γ , is

$$r(D) = \text{Probability of correct classification}$$

$$= \sum_{i=1}^{k} q_i \int_{D_i} f_i(x)\, d\mu(x)$$

(2.2.12) $$= \sum_{i=1}^{k} \int_{D_i} g_i(x)\, d\mu(x) \qquad \text{(see Remark 2.2(ii))}.$$

The rule D was defined to be optimal if it minimized the probability of misclassification. Equivalently, a rule D is optimal if it maximizes the probability of correct classification over the domain $\mathcal{D}^*$ of all classification rules. Let

(2.2.13) $$r^* = \sup_{D \in \mathcal{D}^*} r(D) \quad .$$

Then r^* is called the optimal probability. From the above definition, a classification procedure D is optimal if $r(D) = r^*$. From (2.2.5), the optimal partition D^* is defined by

$$D_j^* = \{x \in \mathcal{X} : g_j(x) = \gamma(x)\} \quad , \quad 1 \leq j \leq k \quad .$$

Now,

$$\begin{aligned} r^* = r(D^*) &= \sum_{j=1}^{k} \int_{D_j^*} g_j(x)\, d\mu(x) \qquad \text{(see (2.2.12))} \\ &= \sum_{j=1}^{k} \int_{D_j^*} \gamma(x)\, d\mu(x) \\ &= \int_{\mathcal{X}} \gamma(x)\, d\mu(x) \quad , \qquad (2.2.14) \end{aligned}$$

which is an expression for the optimal probability of correct classification. For the case of two arbitrary distributions, we have

$$\begin{aligned} \gamma(x) &= \max\{g_1(x), g_2(x)\} \\ &= \frac{1}{2}[g_1(x)+g_2(x)] + \frac{1}{2}|g_1(x)-g_2(x)| \\ &= \frac{1}{2}\{q_1f_1(x)+q_2f_2(x)\} + \frac{1}{2}|q_1f_1(x)-q_2f_2(x)| \quad . \end{aligned}$$

Thus by (2.2.14)

$$\begin{aligned} r^* &= \int_x \frac{1}{2}(q_1f_1(x)+q_2f_2(x))\, d\mu(x) + \frac{1}{2}\int |q_1f_1(x)-(1-q_1)f_2(x)|\, d\mu(x) \\ &= \frac{1}{2} + \frac{1}{2}\int_{\mathcal{X}} |q_1f_1(x)-(1-q_1)f_2(x)|\, d\mu(x) \quad . \end{aligned}$$

In the case of two multivariate normal populations with mean vectors $\mu^{(1)}$ and $\mu^{(2)}$ and common covariance matrix Σ , the simple loss function and equal prior probabilities imply that

$$r^* = \Phi(\frac{\Delta_p}{2}) \quad ,$$

where Δ_p^2 is the Mahalanobis generalized squared distance and Φ is the c.d.f. of standard normal variate.

§2.3 Sample-Based Classification Rules.

In section 2.2, the classification procedures all had an underlying assumption, that the densities have a specified parametric forms, with all parameters known. In most cases, however, the population parameters are usually not known, but must be estimated from the samples. On the basis of information available from the samples, we wish to classify an individual into one of a finite number of populations. It was noted in Section 2.2.5, that the optimal rule, D^* , and $r(D)$, the probability of correct classification for an arbitrary rule $D \in \mathcal{D}^*$, could not be determined unless the distributions F_i , $(i=1,2,\dots,k)$ and the prior probabilities q_i , $(i=1,2,\dots,k)$, were specified. Two questions arise then:

(i) Not knowing an optimal rule, how do we construct a rule from the sample data;

(ii) Given a rule D from the sample data, when are the actual probability $r(D)$ and the optimum probability r^* approximately equal.

These questions have been answered in the following sections:

2.3.1 Plug-in Rules.

Suppose that the dominating measure μ is specified, but $q_i f_i$ $(1 \leq i \leq k)$ are not specified. Suppose further that our inference is based on a well identified random sample of size n drawn from the mixed population Γ, and $n_1, n_2, \ldots, n_k$ are the number of sampled individuals from $\pi_1, \pi_2, \ldots, \pi_k$ respectively. Thus, each of the n_i is a binomial variable with expectation $n\, q_i$ $(i = 1,2,\ldots,k)$. Since the densities, f_i, $1 \leq i \leq k$, involve unknown parameters, the main problem in obtaining "plug-in" rules is to get reasonable estimates of these unknown parameters. Generally, the maximum-likelihood or consistent estimates are used. The corresponding estimates are substituted in place of the unknown parameters to give an estimate of the densities f_i, $1 \leq i \leq k$. Ghurye and Olkin [1969] give parametric multivariate normal density estimates that are pointwise unbiased.

If we have estimates $\widehat{q_i f_i}$, $1 \leq i \leq k$, then evidently an intuitive choice of rule is that rule $\hat{D}$ obtained by substituting $\widehat{q_i f_i}$ for $q_i f_i$ in the expression (2.2.5) for the optimal rule D^*. Similarly, we can substitute the estimates into the expression (2.2.12) for $r(D)$. We call $\hat{D}$ the "plug-in" rule. In most instances, we use the estimates

$$\widehat{q_i f_i}(x) = \hat{q}_i \hat{f}_i(x) \quad , \quad 1 \leq i \leq k$$

where

(2.3.1) $$\hat{q}_i = \frac{n_i}{n} \quad , \quad 1 \leq i \leq k$$

and $\hat{f}_i$ is some estimate of the density f_i $(1 \leq i \leq k)$ obtained by substituting the estimates for the unknown parameters.

The estimates $\hat{q}_i$ given by (2.3.1) are quite well behaved. They satisfy

(2.3.2) $$\hat{q}_i \xrightarrow{a\cdot s} q_i \quad \text{as} \quad n \to \infty$$

by the strong law of large numbers. If these q_i's are known, then one uses $\widehat{q_i f_i}(x) = q_i \hat{f}_i(x)$, $1 \leq i \leq k$. One also obtains immediately the estimates

$$\hat{g}_j(x) = \hat{q}_j \hat{f}_j(x) \quad , \quad 1 \leq j \leq k$$

$$\hat{\gamma}(x) = \max_{1 \leq j \leq k} \hat{g}_j(x)$$

of $g_j(x)$ and $\gamma(x)$ respectively.

Throughout the classification literature, the plug-in rules seem to be the only rule choices ever considered when specifications are incomplete. The general theory has not yet been studied satisfactorily. All one can do is to substitute the estimates of unknown parameters. In case of plug-in rules, the optimality criterion can no longer be justified except for large samples for which the performance of the plug-in rule $\hat{D}$ is, in some sense, close to that of the optimal rule D^* . In fact, due to sampling variations in the estimation of the parameters, the plug-in rule $\hat{D}$ is no longer the best. The only justification

Anderson [1958] gives for the use of plug-in linear discriminant is that, "it seems intuitively reasonable that this rule should give good results". The following are some special cases:

(i) Anderson's Rule:

Suppose we have samples $x_1^{(1)},\ldots,x_{n_1}^{(1)}$; and $x_1^{(2)},x_2^{(2)},\ldots,x_{n_2}^{(2)}$; from two multivariate normal populations π_1 and π_2 respectively, with all parameters $\mu^{(1)}$, $\mu^{(2)}$ and the common covariance matrix, Σ , unknown. In the case of known parameters, the optimal rule D^* was defined by

$$D_1^* = \{X \in \mathcal{X} : X'\Sigma^{-1}(\mu^{(1)}-\mu^{(2)}) - \tfrac{1}{2}(\mu^{(1)}+\mu^{(2)})'\,\Sigma^{-1}(\mu^{(1)}-\mu^{(2)}) > \log C\}$$

$$D_2^* = \mathcal{X} - D_1^* \ .$$

Since, in this case, the parameters are unspecified, the usual plug-in linear discriminant is that rule $\hat{D}$, obtained by substituting the best (namely unbiased) estimates of these unknown parameters. Consequently, the plug-in rule, $\hat{D}$, is given by

$$(2.3.3)\quad \begin{cases} \hat{D}_1 = \{X \in \mathcal{X} : X'\,S^{-1}(\bar{x}^{(1)}-\bar{x}^{(2)}) \\ \qquad - \frac{1}{2}\,(\bar{x}^{(1)}+\bar{x}^{(2)})'\,S^{-1}(\bar{x}^{(1)}-\bar{x}^{(2)}) > \log C\} \\ \\ \hat{D}_2 = \mathcal{X} - \hat{D}_1 \ . \end{cases}$$

The term $X' S^{-1}(\bar{x}^{(1)}-\bar{x}^{(2)})$ is the linear discriminant based on two samples and is called Anderson's plug-in linear discriminant. The classification statistic is denoted by $V(x)$; i.e.,

$$V(x) = X' S^{-1}(\bar{x}^{(1)}-\bar{x}^{(2)}) - \frac{1}{2}(\bar{x}^{(1)}+\bar{x}^{(2)})' S^{-1}(\bar{x}^{(1)}-\bar{x}^{(2)}) \quad . \tag{2.3.4}$$

Anderson [1958] has obtained the asymptotic distribution of V . Its exact distribution is not known explicitly. He has shown that its limiting distribution approaches the distribution of U ((2.2.8)) as the sample sizes increase indefinitely. Hence, for sufficiently large samples from π_1 and π_2 we can proceed as if the parameters were completely specified.

(For an example of this rule of classification, see Appendix I.)

(ii) Mahalanobis' Studentized D_p^2 :

The plug-in version of Mahalanobis' generalized squared distance, Δ_p^2 , is his studentized D_p^2 , obtained by replacing the unknown parameters $\mu^{(i)}$, $\mu^{(j)}$, and Σ by their corresponding 'best' estimates. Let there be two samples of sizes n_1 and n_2 from π_1 and π_2 respectively. D_p^2 is given by

$$D_p^2 = \sum_i \sum_j s^{ij} d_i d_j \tag{2.3.5}$$

where d_i denotes the difference in the mean values for the ith variable in the two samples; (s^{ij}) denotes the elements of the inverse matrix of the estimate of the common or pooled covariance matrix.

Putting (2.3.5) in the matrix notation, we get

(2.3.6) $$D_p^2 = (\bar{x}^{(i)} - \bar{x}^{(j)})' \, S^{-1} (\bar{x}^{(i)} - \bar{x}^{(j)}) \quad .$$

Consequently, mimicking what was done in section 2.2.4, the plug-in minimum distance rule classifies an observation X_o into π_1 or π_2 according as

(2.3.7) $$(X_o - \bar{x}^{(1)})' \, S^{-1} (X_o - \bar{x}^{(1)}) \lessgtr (X_o - \bar{x}^{(2)})' \, S^{-1} (X_o - \bar{x}^{(2)}) \quad .$$

An increase in D_p^2 due to the additional information supplied by new variables is not appreciable. A higher value of the ratio

$$R = \frac{1 + \frac{n_1 n_2}{(n_1+n_2)(n_1+n_2-2)} D^2_{p+q}}{1 + \frac{n_1 n_2}{(n_1+n_2)(n_1+n_2-2)} D_p^2} \quad ,$$

would indicate that q new variables supply some information (see Rao [1952].)

(For an example of this result, see Appendix I.)

Result 2.4: D_p^2, the Mahalanobis' studentized distance, is not an unbiased estimate of Δ_p^2, the Mahalanobis' generalized squared distance.

2.3.2 Likelihood-Ratio Criterion.

Another criterion that could be considered in the classification theory is the likelihood-ratio criterion, first introduced by

Anderson [1951]. Let the class densities be known except for some parameters. For example the populations may have multivariate normal densities with common unknown covariance matrix and unknown mean vectors.

Let n be the size of the "training" sample and n_i be the size of the sample from π_i $(i = 1,2,\ldots,k)$. Let n_o be the size of the sample from π_o , which is to be classified. We shall denote such a sample by "CS". Let $L(TS)$ denote the likelihood of the "training" sample and $L_i(CS)$ denote the likelihood of CS under the hypothesis $\pi_o = \pi_i$, $i = 1,2,\ldots,k$.

Let

$$\lambda_i = \sup \left\{\frac{L_i(CS)}{L(TS)}\right\} \quad ,$$

the supremum being taken over the parametric space.

A <u>likelihood-ratio</u> <u>rule</u> (LR rule) classifies CS into π_i iff

$$k_i \lambda_i = \max_{1\leq j\leq k} (k_j\lambda_j)$$

where k_i's are non-negative constants. Ties may be resolved in some manner.

A maximum-likelihood rule (ML rule) is a LR rule with equal k_i's . Equivalently, a ML rule classifies an observation X_o into π_i if ML obtained under the assumption that X_o comes from π_i is greater than the corresponding ML assuming that the observation X_o comes from π_j , $j \neq i$.

As a particular case, consider the classification of an observation, X_o , into one of two multivariate normal populations, π_1 and π_2 , with all parameters unknown. Let $x_1^{(1)},\ldots,x_{n_1}^{(1)}$ and $x_1^{(2)},x_2^{(2)},\ldots,x_{n_2}^{(2)}$ be the samples of sizes n_1 and n_2 from π_1 and π_2 respectively. Considering the maximum likelihood estimates of the parameters under the two hypotheses that X_o comes from π_1 , and X_o comes from π_2 , the ML rule classifies an observation, X_o , into π_1 or π_2 according as

(2.3.8)
$$(1+n_1^{-1})^{-1}(X_o-\bar{x}^{(1)})'\ S^{-1}(X_o-\bar{x}^{(1)})$$
$$\lessgtr (1+n_2^{-1})^{-1}(X_o-\bar{x}^{(2)})'\ S^{-1}(X_o-\bar{x}^{(2)}) + A$$

If Σ is known, then S is replaced by Σ in the above expression. (For details see Anderson [1958] pp. 141.)

Das Gupta [1965] considers the above ML rule and has established that it is an unbiased, admissible minimax rule. Further, if the loss function ℓ is continuous such that

(2.3.9)
$$\lim_{y\to 0} \ell(y) = 0$$

then ML rule is the unique minimax rule. In case of unknown Σ , Das Gupta proves that the ML rule is unbiased, admissible minimax in an invariant class and if the loss function ℓ is continuous satisfying (2.3.9), then it is the unique minimax rule in the invariant class.

Remark 2.5: In the case of classification into one of two multivariate normal populations π_1 and π_2 with parameters unspecified, the MD rule, the ML rule, and Anderson's rule are special cases of the following rule:

Classify an observation, X_o, into π_1 or π_2 according as

$$(2.3.10)\quad a(X_o-\bar{x}^{(1)})'\, S^{-1}(X_o-\bar{x}^{(1)}) \lessgtr (X_o-\bar{x}^{(2)})'\, S^{-1}(X_o-\bar{x}^{(2)}) + b \quad .$$

For example:

(i) $a = (1+n_1^{-1})^{-1}(1+n_2^{-1})$ and $b = (1+n_2^{-1})A$, gives the ML rule.

(ii) $a = 1$ and $b = 0$ gives the MD rule.

(iii) $a = 1$ and $b = -2 \log C$ gives Anderson's rule.

Remark 2.6: We have so far considered procedures for classifying an individual into one of many populations, specified completely or not, with quantitative data. Sometimes however, the data is qualitative or "categoric" (known only by its category). In that case, the variables have discrete distributions. The most familiar instance is the process of medical diagnosis using laboratory tests with discrete outcome states, -/+ ; -/?/+ ; or milky/greenish/clear/dark etc. (for a liquid). Glick [1973] considers this problem at length and arrives at sample-based multinomial classification rules. He has also obtained some results on the asymptotic optimality of these rules. For a detailed discussion on this topic, one is referred to Glick [1969, 1973].

2.3.3. On the Estimation of the Probability of Correct Classification.

There are at least two reasons for wanting to know the probability of correct classification, of a classification procedure. One is to see if the classification rule performs well enough to be useful. Another is to compare its performance with a competing rule. In sections 2.2.2 and 2.2.5, we obtained an expression for the optimal rule, D^*, and for r^*, the optimal probability of correct classification, respectively, when the distributions were completely specified. In the case of unspecified parameters, section 2.3.1 discusses the choice of plug-in rules, $\hat{D}$, obtained by using suitable estimates of the unknown parameters in the expression for D^*. Corresponding to this $\hat{D}$, $r(\hat{D})$ denotes the probability of correct classification.

The density plug-in estimator, $\hat{r}$, of the optimum probability $r^* = r(D^*)$, is defined by

$$\hat{r} = r(\widehat{D^*}) = \sum_{i=1}^{k} \int_{\hat{D}_i} \hat{q}_i \hat{f}_i \tag{2.3.11}$$

where

$$\hat{D}_i = \{X \in \mathcal{X} : \hat{g}_i(x) = \hat{\gamma}(x)\} \quad , \quad 1 \leq i \leq k$$

are components of the partition of $\hat{D}$. The probability of correct classification for the plug-in rule $\hat{D}$ has the expression,

$$r(\hat{D}) = \sum_{i=1}^{k} \int_{\hat{D}_i} q_i f_i(x)\, d\mu(x) \quad . \tag{2.3.12}$$

(This is analogous to (2.2.12), for $r(D)$.)

Mimicking what we did to arrive at (2.2.13) we get,

$$\hat{r} = \int_{\mathcal{X}} \hat{\gamma}(x)\, d\mu(x) \quad .$$

In (2.3.12), if we substitute the estimates of $q_i f_i(x)$, we get $\hat{r}$ as an estimate of $r(\hat{D})$ as well. Thus,

$$\hat{r} = \hat{r}(\hat{D}) = \int_{\mathcal{X}} \hat{\gamma}(x)\, d\mu(x) \quad . \tag{2.3.13}$$

Thus, the plug-in approach yields the same estimator, $\hat{r}$, as an estimate of both the optimal probability, r^*, and the actual probability of correct classification for $\hat{D}$, $r(\hat{D})$. Glick [1972] has shown that if the estimates $\widehat{q_i f_i}$ are pointwise unbiased, or more generally satisfy

$$E(\hat{q}_i \hat{f}_i(x)) \geq q_i f_i(x) \quad , \quad 1 \leq i \leq k \; , \quad \text{almost all} \quad x \in \mathcal{X} \quad ,$$

then $\hat{r}$ is biased as an estimate of either the optimal probability or the actual probability of correct classification, $r(\hat{D})$. (For proof see theorem 4.1.) Glick also states general conditions under which $\hat{r}$ is a consistent estimate of r^*. (Theorems 4.2, 4.3 and 4.4 - for proofs see section 4.1 of Chapter IV.)

Lachenbruch and Mickey [1968] have suggested a number of methods for estimating the two components of the probability of misclassification, namely

$$P_1 = P\{V(X) < 0 \mid X \in \pi_1\}$$

and

$$P_2 = P\{V(X) > 0 \mid X \in \pi_2\}$$

where $V(X)$ is Anderson's statistic given in (2.3.4). The techniques may be divided into two classes: those using a sample to evaluate a given discriminant function and those using the properties of normal distribution. The second approach depends heavily on the normality for their validity. For the multivariate case, Lachenbruch and Mickey [1968] comment that their "method D" tends to be "badly biased and give much too favourable an impression of the probability of error". They studied a comparative evaluation of all their suggested methods of estimation of P_1 and P_2 on the basis of a series of Monte Carlo experiments. They concluded that no one method is uniformly best for every situation, although D and R methods appear to be relatively poor and the O method does fairly well. (For a discussion of these methods, see Lachenbruch and Mickey [1968] or Kshirsagar [1972].)

Remark 2.7: In case of a sample-based classification procedure classifying an individual into one of two multivariate normal populations, the probability of correct classification is not $\Phi(\frac{\Delta_p}{2})$, nor can it be obtained in a similar manner.

2.3.4 Step-Down Procedure.

In most classification procedures, it would be desirable to find the magnitude of the errors committed. Consequently, much of the attention is devoted towards obtaining the exact and asymptotic distributions of the classification statistics. In most cases, the usual

asymptotic expression for error is an underestimate of the actual error (see Srivastava [1973]). Srivastava [1973] proposes the "step-down" procedure when the variates can be arranged according to their importance on a priori grounds.

Let the two populations have multivariate normal densities with the same covariance matrix, Σ . The classification is carried out on the basis of the marginal univariate distribution of the first variate, on the conditional univariate distribution of the second variate given the first, on the conditional univariate distribution of the third variable given the first and the second, and so on. Let

$X' = [x_1,\ldots,x_p]$ be the vector to be classified.

$X'_{(i)} = [x_1,x_2,\ldots,x_i]$.

We define $Y'_{(i)},Z'_{(i)},\mu^{(j)}_{(i)}$ $(j=1,2)$ similarly for the two populations. Let the top left-hand $i\times i$ submatrix of $\Sigma = [(\sigma_{ij})]$, be denoted by Σ_i . Let

$$\beta_i = \Sigma_i^{-1} \begin{bmatrix} \sigma_{1,i+1} \\ \sigma_{2,i+1} \\ \cdot \\ \cdot \\ \cdot \\ \sigma_{i,i+1} \end{bmatrix} \quad , \quad i = 1,2,\ldots,p$$

and $\sigma^2_{i+1} = \dfrac{|\Sigma_{i+1}|}{|\Sigma_i|}$, $i = 0,1,2,\ldots,p-1$ with the convention that $\beta_o = 0$ and $|\Sigma_o| = 1$ so that $\sigma_1{}^2 = \sigma_{11}$. We call β_i , the ith

order step-down regression coefficient and σ^2_{i+1} , the ith order step-down residual variance. Let

$$\eta^{(j)}_{i+1} = \mu^{(j)}_{i+1} - \mu^{(j)'}_{(i)} \beta_i \ , \qquad \begin{array}{l} i = 0,1,2,\ldots,p-1 \\ j = 1,2 \ . \end{array}$$

Then under the condition that $Y_{(i)}$ is fixed, the conditional distribution of Y_{i+1} is normal with mean $\eta^{(1)}_{i+1} + Y'_{(i)} \beta_i$ and variance σ^2_{i+1} . The distributions of z_{i+1} given $Z_{(i)}$ and x_{i+1} given $X_{(i)}$ are similar.

Let $\hat{\beta}$ be the usual (replacing the unknown parameters by 'best' sample estimates) estimator of β . Let, for $i = 0,1,2,\ldots,p-1$,

$$(2.3.14) \qquad \begin{cases} \tilde{x}_{i+1} = x_{i+1} - X'_{(i)} \hat{\beta}_i \\ \tilde{y}_{i+1} = y_{i+1} - Y'_{(i)} \hat{\beta}_i \\ \tilde{z}_{i+1} = z_{i+1} - Z'_{(i)} \hat{\beta}_i \ . \end{cases}$$

Then the step-down procedure classifies an individual with measurements $\underset{\sim}{X}$ into π_1 if for all $i = 1,2,\ldots,p$

$$(2.3.15) \qquad Q_i \equiv \tilde{x}_i(\tilde{y}_i - \tilde{z}_i) - \frac{1}{2}(\tilde{y}_i + \tilde{z}_i)(\tilde{y}_i - \tilde{z}_i) > 0 \ ,$$

and to π_2 if for all $i = 1,2,\ldots,p$, $Q_i < 0$; otherwise it is assigned to neither π_1 nor π_2 . (For an expression for probability of misclassification for this procedure, see Srivastava [1973].)

Remark 2.8: In the step-down procedure, an individual may not be classified at all to any of the two populations π_1 , π_2 . This is one

of the features of this procedure, for it is better not to assign to any one of the two in the absence of sufficient evidence. The procedure is clearly not invariant under permutation of the variates, and should be used only when the variates can be arranged on a priori grounds.

CHAPTER III

Nonparametric Classification

In Chapter II, we discussed some major parametric rules of classification and the associated probabilities of misclassification. These techniques assume the existence and knowledge of the underlying probability densities. In practice however, the forms of the underlying distributions are seldom known and one is often confronted with the problem of devising appropriate classification rules, applicable for a wider class of distributions, whose structures are not expressible in simple parametric forms. In such cases, the use of parametric procedures is subject to criticism regarding its appropriateness and validity. For such situations, one uses the so-called "nonparametric" or "distribution-free" methods, which are the subject of this chapter.

§3.1 Statement of the Problem.

The problem is to classify units into a specified number of populations on the basis of a set of observations on these units, with all population distributions F_i's $(i=1,2,\ldots,k)$ unspecified. Some assumptions, however, are needed for constructing discriminant rules, for example, the existence of densities, a unique mode etc. In case of nonparametric classification procedures, the main emphasis is:

(i) to study the asymptotic behaviour of the rules (e.g., consistency, efficiency),

(ii) to obtain suitable bounds for the probability of correct classification.

§3.2 On the Estimation of the Probability Density Function.

A basic and important problem in nonparametric classification techniques is the estimation of the assumed probability density function and its mode. Discriminant criteria are then based on the estimates of these assumed densities. There are two forms of density estimation - parametric and nonparametric.

3.2.1 Nonparametric Density Estimation.

If the functional form of the density is known but depends upon a finite number of unknown parameters, the usual method of estimation would be to obtain suitable estimates of these unknown parameters and plug-in these estimates in place of unknown parameters giving an estimate of the parametrized density. This case was discussed in Chapter II, to obtain the so-called "plug-in" rules of classification.

Let $f_1, f_2, \ldots, f_k$ be the densities with respect to a σ-finite measure μ. Fix and Hodges [1951] were the first who considered nonparametric density estimation in connection with nonparametric discrimination. Parzen [1962] and later Cacoullos [1966], who generalized Parzen's work to the multivariate case, developed a class of nonparametric density estimates having the form

$$\hat{f}_i(x) = \frac{1}{h}\int_{-\infty}^{\infty} k\left(\frac{x-y}{h}\right) d\,\hat{F}_i(y) \quad , \quad 1 \leq i \leq k$$

$$(3.2.1) \qquad = \frac{1}{n_i h}\sum_{j=1}^{n_i} k\left(\frac{x-X_{ij}}{h}\right) \quad , \quad 1 \leq i \leq k$$

where X_{ij} is the jth - sample observation from π_i , $\hat{F}_i$ is the empirical distribution function of the n_i individuals sampled from π_i $(i = 1,2,\ldots,k)$, $k(x)$ is a bounded Lebesgue integrable function on $(-\infty,\infty)$ such that

$$\lim_{x\to\infty} |x\ k(x)| = 0$$

(3.2.2) $$\int_{-\infty}^{\infty} k(x)\ dx = 1 \quad ,$$

and $h = h(n)$, where $n = \sum_{i=1}^{k} n_i$, is a non-negative sequence satisfying

(3.2.3) $$\lim_{n\to\infty} h(n) = 0 \quad .$$

Functions $k(x)$ of the above type satisfying (3.2.2) are called 'weighting' or 'Kernel' functions. It should be noted that the choice of $k(x)$ is very important, and to a large extent determines the properties of $\hat{f}_i(x)$. One simple example of a kernel function is

$$k(x) = \begin{cases} \frac{1}{2} & |x| \leq 1 \\ 0 & |x| > 1 \end{cases} \quad .$$

For different choices of kernel functions, see table 1 of Parzen [1962].

This definition includes the special cases of the form

$$\hat{f}_i(x) = \frac{\hat{F}_i(x+h) - \hat{F}_i(x-h)}{|h|} \quad , \quad 1 \leq i \leq k$$

where $|h| \to 0$ as $n \to \infty$. The estimates suggested by Parzen-Cacoullos are also called "Fixed window" estimates. If, in addition to (3.2.3), $h = h(n)$ satisfies

$$\lim_{n\to\infty} n\, h(n) = \infty \quad , \tag{3.2.4}$$

then Parzen [1962] proved that these density estimates are consistent. He also proved the asymptotic normality of the estimates. (For details see Parzen [1962].) Using Parzen's density estimates (3.2.1), and added conditions, Glick [1969, Theorem 6d, pp. 72] proves the consistency of the plug-in estimator $\hat{r}$ of the optimum probability r^* .

Suppose D^* is a Bayes rule with respect to a prior distribution, assuming the densities in the k populations are known. Let $\hat{D}$ be the plug-in rule. By remark 2.2(i) $\rho(D^*)$ denotes the Bayes risk of the optimal rule D^* . Let $R(\hat{D})$ denote the Bayes risk of the plug-in rule $\hat{D}$. Van Ryzin [1966] introduces the notion of "Bayes risk consistency", defined by the following.

<u>Definition 3.1</u>. The rule $\hat{D}$ is Bayes risk consistent (BRC) with D^* if

$$P[R(\hat{D}) - \rho(D^*) \geq \epsilon] \to 0$$

as the sample sizes in the training sample tend to ∞ .

With respect to this notion, using Parzen-Cacoullos density estimates, Van Ryzin [1966] studied the asymptotic optimality of sample-based classification rules. For related results see Van Ryzin [1965] and

section 4.2 of Chapter IV. Van Ryzin [1969] gives conditions for the pointwise 'almost sure' convergence of the fixed window estimates. ("Potential functions" in the Pattern recognition theory are synonyms for the "Kernel" of the fixed window density estimation theory.)

An alternative nonparametric approach for estimating multivariate densities has been proposed by Loftsgaarden and Quesenberry [1965]. Let

$$S_d(x) = \{y \in \mathcal{X} : ||y-x|| \leq d\}$$

and denote the volume of this hypersphere by

$$A_{d,x} = \mu(S_d(x)) \quad .$$

Let k_n be a non-decreasing sequence of positive integers such that $k_n \to \infty$, but $\frac{k_n}{n} \to 0$ as $n \to \infty$. Let $d_{k_{n_i}}(x)$ be the distance from X to the k_{n_i}-th closest point among n_i sampled individuals from the density f_i $(i = 1,2,\ldots,k)$. Then the Loftsgaarden and Quesenberry estimate of $f_i(x)$ is

$$(3.2.5) \qquad \hat{f}_i(x) = \frac{k_{n_i} - 1}{n_i \, A_{d_{k_{n_i}}(X),X}} \quad , \qquad i = 1,2,\ldots,k \quad .$$

In contrast to the fixed window estimates, these estimates given by (3.2.5) are called "variable window" or "fixed view". Glick [1969] proved that if μ is a Lebesgue measure and $\widehat{q_i \, f_i}(x) = \frac{n_i}{n} \hat{f}_i(x)$, where each $\hat{f}_i$, $1 \leq i \leq k$, is of the form (3.2.5), then the plug-in

estimator $\hat{r}$ is a consistent estimate of the optimum probability r^*. (For proof see Theorem 4.2.)

In fact, there are several other papers in the literature dealing with various methods for estimating probability density functions and their properties. For more detailed references in this connection, see Cacoullos [1973], Glick [1972] and Patrick [1972]. In particular, Patrick [1972] gives an excellent account of estimation by the potential functions methods and stochastic approximation method - a method of searching for a parameter vector which optimizes a prescribed criterion. A final remark on nonparametric density estimation: For univariate unimodal densities, B.L.S.P. Rao [1969] shows that the maximum likelihood density estimate is "the slope of the concave majorant of the empirical distribution" and that this estimate, too, is consistent (converges pointwise in probability). (Maximum likelihood density estimation can also be found in Wegman [1970a, 1970b]. In general, as remarked by Wegman [1972] the maximum likelihood density estimates do not exist, but with some appropriate type of restriction on the class of densities from which the density may be selected a maximum likelihood estimate over that class may exist.)

§3.3 <u>Classification Rules</u>.

3.3.1 <u>Nearest Neighbor Rule</u>.

Throughout this section, simple loss structure, namely,

$$C_{ij} = \begin{cases} 0 & i = j \\ 1 & i \neq j \end{cases} \qquad \text{(see (2.2.1))}$$

is assumed and formulation (ii) of section 2.1, with q_i's , $1 \leq i \leq k$, unknown is considered.

In section 2.3.3, we considered density plug-in estimator $\hat{r}$ of the optimum probability of correct classification r^* . Three problems arise in this estimation:

(i) Such an estimate is almost always over optimistic,

(ii) one should always suspect the validity of an assumed parametric model,

(iii) in more general situations it is quite difficult to compute these probabilities exactly, even if the probabilistic stucture is completely known.

In order to overcome some of these drawbacks Glick [1969] introduces the notion 'counting' estimate of the probability of correct classification, and gives a classification rule related to this notion. Let D be any classification procedure, namely, an ordered-partition $<D_1, D_2, \ldots, D_k>$ of the sample space $\mathcal{X}$. Given a correctly classified random sample of size n from the mixed population Γ , the proportion of sampled individuals who would be correctly classified by D is the most natural estimate of the rule's actual probability of correct classification. This estimate is known as counting or empiric estimate and is denoted by $\tilde{r}(D)$. Thus,

$$\tilde{r}(D) = \frac{1}{n} \text{ (\# of sampled individuals classified correctly by D)}$$

$$= \frac{1}{n} \text{ (\# of } \pi_i \text{ sampled individuals classified correctly by D)}$$

$$= \frac{1}{n} \left(\# \text{ of } \pi_i \text{ sampled individuals with } X \in D_i\right)$$

$$= \frac{1}{n} \sum_{i=1}^{k} \int_{D_i} d(n_i \hat{F}_i(x))$$

$$\text{(3.3.1)} \qquad = \sum_{i=1}^{k} \int_{D_i} \hat{q}_i \, d \, \hat{F}_i(x) \qquad \text{(see (2.3.1))}.$$

The counting function $\tilde{r}$ resembles the density plug-in estimate $\hat{r}$, but the empiric approach differs in an important way from the plug-in approach, viz, no restriction is placed on the distributions F_i $(i = 1,2,\ldots,k)$, the densities $f_i = \frac{dF_i}{d\mu}$ are of no importance and μ need not be specified. For these reasons the nearest neighbor rule, to be described below, is termed a nonparametric classification procedure. Also note that, unlike $\hat{r}$, the counting estimate $\tilde{r}$ is an unbiased estimate of $r(D)$ for:

$$E(\tilde{r}(D)) = \frac{1}{n} \sum_{i=1}^{k} E\,(\# \text{ of sampled } \pi_i \text{ individuals with } X \in D_i\,)$$

$$= \frac{1}{n} \sum_{i=1}^{k} q_i \, n \, P\,(X \in D_i \,/\, \underset{\sim}{X} \text{ is drawn from } \pi_i\,)$$

$$= \sum_{i=1}^{k} q_i \int_{D_i} d\, F_i(x)$$

$$\text{(3.3.2)} \qquad = r(D) \qquad \text{(see (2.2.12))}.$$

Mimicking the optimality criterion, one would desire to have a classification procedure that maximizes the counting function $\tilde{r}$. Since $\tilde{r}(D) \leq 1$ for any discriminant D, consequently $\sup_{D \in \mathcal{D}^*} \tilde{r}(D) \leq 1$ a·s

The equality $\sup_{D\in\mathcal{D}^*} \tilde{r}(D) = 1$ is attained by the so called <u>nearest neighbor rule</u> (NN rule) $\ddot{D}$, which assigns an unidentified individual from the mixed population Γ to the category of a nearest correctly classified sample observation.

<u>Definition 3.2</u>: We call $x_n \in \{x_1,x_2,\ldots,x_n\}$, a nearest neighbor to x , if

$$\min_{1\leq i\leq n} d(x_i,x) = d(x_n,x) \quad .$$

The distance, in general, may be other than the usual euclidean distance.

The NN rules are ones among a broad category of "good" data dependent rules, distinct from the plug-in rule $\hat{D}$ discussed in section 2.3.1. The first formulation of the NN rule and contribution to the analysis of its properties were made by Fix and Hodges, as early as 1951. Subsequently, these rules have been investigated by Cover and Hart [1967] and Cover [1968]. Variations on this theme include the ν - nearest neighbor rule, which assigns an unidentified individual to a subpopulation with a plurality among the ν measurements. Cover and Hart [1967] have shown among the class of all ν - nearest neighbor rules, the simple nearest neighbor rule is admissible. They prove the convergence $r(\ddot{D}) \to r_{NN}$ with probability one, and for $k = 2$ classes, the limit is bounded by (see Cover and Hart [1967]),

$$\begin{aligned} 1 \geq r^* &\geq r_{NN} \\ &\geq 1-2r^*(1-r^*) \geq \frac{1}{2} \end{aligned}$$

and $r^* = r_{NN}$ iff $r^* = \frac{1}{2}$ or $r^* = 1$, i.e. in the two extreme cases of complete certainity and complete uncertainity, the nearest neighbor actual probability of correct classification equals the optimum probability. It is in these cases, or approximations to it, that the nearest neighbor rule is most useful. Later in 1968, Cover [1968] studied the rate of convergence of the Bayes risk of their nearest neighbor rule. An excellent account of nearest neighbor rules is given in Patrick [1972].

Finally, we must specify means of resolving the tie, for example, the rule may be modified to decide the most popular category among the ties or assigning to that population with lowest subscript. Glick [1969] remarks that "A NN rule seems most reasonable and useful when the probability of ties is zero". However, Cover and Hart [1967] claim that their results are true even for those cases in which the ties occur with non-zero probability. This assertion, however, seems to need some mathematical justification.

Remark 3.1: If the probability of ties is zero, then with probability one, the rule $\ddot{D}$ classifies correctly all n sampled individuals, i.e. $\tilde{r}(\ddot{D}) = 1$. Hence, Glick [1969] comments that "the counting estimate of $r(\ddot{D})$, the simple NN rule's probability of correct classification is grossly biased and unreasonable". Due to this fact, further methods of estimation of $r(D)$ are suggested in the literature. One of such methods is deletion-counting method of estimation, which is not dealt with here. One is referred to Glick [1969] for further details.

(For an example of the NN rule, see Appendix I.)

3.3.2 Minimum Distance Classification Rule.

Das Gupta [1964] suggested the so-called minimum distance classification rule for the above nonparametric classification problem. Let $X(p\times 1)$ be a random vector from one of the populations π_i $(i = 0,1,2,\ldots,k)$ with distribution functions F_i $(i = 0,1,2,\ldots,k)$. The F_i's are completely unspecified except that $F_o = F_i$ for exactly one value of i , $(i = 1,2,\ldots,k)$ and F_i's $(i = 1,2,\ldots,k)$ are all distinct. Let D denote the decision space $(d_1,\ldots,d_k)$ where d_i denotes the decision $F_o = F_i$, $i = 1,2,\ldots,k$. Let $\underset{\sim}{X}$ be a vector of sample observations. Then a classification rule $\phi = (\phi_1,\phi_2,\ldots,\phi_k)$ is a k - dimensional vector valued measurable function of $\underset{\sim}{X}$ such that

$$(3.3.3) \qquad \left.\begin{array}{l} 0 \leq \phi_i(\underset{\sim}{X}) \leq 1 \\ \\ \sum_{i=1}^{k} \phi_i(\underset{\sim}{X}) = 1 \end{array}\right\} \quad \forall\ \underset{\sim}{X} \in \mathcal{X}$$

and $\phi_i(\underset{\sim}{X})$ denotes the probability of taking the decision d_i on observing $X = x$.

Definition 3.3: The minimum distance rule, $\phi^{(d)}$, based on a p - variate distance function d (arbitrary distance), is defined by

$$(3.3.4) \qquad \phi_i^{(d)}(\underset{\sim}{X}) = \begin{cases} 1 & \text{if } d_{oi} = \min\limits_{1\leq j\leq k} doj \\ \\ 0 & \text{otherwise} \end{cases}$$

for $i = 1,2,\ldots,k$, where $d_{oi} = d(\hat{F}_o,\hat{F}_i)$, $1 \leq i \leq k$, $\hat{F}_o$ being the empirical distribution function of n_o individuals sampled from π_o .

Definition 3.4: A distance function d between two p - variate distribution functions is said to be consistent if, given any $\epsilon > 0$, $\epsilon' > 0$, there exists a number N such that for $n > N$

$$P[d(\hat{F}_n, F) \geq \epsilon | F] < \epsilon' \tag{3.3.5}$$

where $\hat{F}_n$ is the sample distribution function, based on a random sample of size n from a p - variate population with distribution function F.

If (3.3.5) holds uniformly for all $F \in B$, a subclass of all p - variate distribution functions, then D is said to be uniformly consistent (B).

Definition 3.5: A distance function d is called the Kolmogorov-distance when

$$d(F,G) = \sup_{-\infty<x<\infty} |F(x) - G(x)| \quad . \tag{3.3.6}$$

Following the definition of $\phi_i^{(d)}(X)$, in (3.3.4), let

$$r_{ii}(d) = P[\phi_i^{(d)}(X) = 1 \mid F_o = F_i] \quad , \quad i = 1,2,\ldots,k \tag{3.3.7}$$

and let

$$f_d(n,\gamma,F) = P[d(\hat{F}_n,F) < \gamma | F] \quad . \tag{3.3.8}$$

With respect to the consistency (uniform) notion of a distance function d, Das Gupta [1964] has proved that the minimum distance classification rule $\phi^{(d)}$ defined by (3.3.4) is consistent (uniform);

i.e., $r_{ii}(d) \to 1$ as $n_i \to \infty$, $i = 1,2,\ldots,k$ if the distance function d is consistent (uniform). He further extends the result to the case when d is the Kolmogorov-distance defined by (3.6). Das Gupta [1964] obtained a lower bound for the probability of correct classification for such rules given by:

$$(3.3.9) \qquad r_{ii}(d) \geq f_d(n_1,\tfrac{\beta}{4},F_1)\; f_d(n_2,\tfrac{\beta}{4},F_2)\; f_d(n_o,\tfrac{\beta}{4},F_o=F_i) \quad (i=1,2) ,$$

where $d(F_1,F_2) > \beta > 0$, and when d is the Kolmogorov-distance

$$(3.3.10) \qquad r_{ii}(d) \geq \prod_{i=0}^{2} \{1 - \frac{16}{\ell_{12}} e^{-n_i \ell_{12}^2/32}\} , \quad i = 1,2 ,$$

where $\ell_{12} = d(F_1,F_2)$. (For proofs of these assertions, see section 4.3 of Chapter IV.)

3.3.3 Classification Rules Based on Ranks.

The idea of using the rank-statistics for devising classification procedures was suggested by Das Gupta [1964]. He proposed the following rule based on the Wilcoxon-Statistic for the classification of an individual into one of two univariate populations.

As in section 3.3.2, let X be a random vector from a population π_o with distribution function F_o , which is one of the two populations π_i $(i = 1,2)$ with continuous distribution functions F_i $(i = 1,2)$ respectively. The properties of the Wilcoxon-Statistic for the discrete case have not been fully investigated so far. Let $(x_1,x_2,\ldots,x_{n_o})$, $(y_1,y_2,\ldots,y_{n_1})$, $(z_1,z_2,\ldots,z_{n_2})$ be random samples of sizes n_o , n_1 , n_2 from populations π_o , π_1 , π_2 respectively.

Define

$$u = \frac{1}{n_o n_1} \text{ X } \# \text{ of pairs } (x_i, y_j) \text{ with } x_i < y_j \quad ,$$

$$(i = 1,2,\ldots,n_o \; ; \; j = 1,2,\ldots,n_1)$$

$$v = \frac{1}{n_o n_2} \text{ X } \# \text{ of pairs } (x_i, z_k) \text{ with } x_i < z_k \quad ,$$

$$(i = 1,2,\ldots,n_o \; ; \; k = 1,2,\ldots,n_2) \quad .$$

The proposed classification rule, based on these statistics u and v, is defined by: Decide

(3.3.11) $$F_o = F_1 \quad \text{if} \quad |u - \tfrac{1}{2}| < |v - \tfrac{1}{2}|$$

decide $F_o = F_2$ otherwise. (3.3.11) is equivalent to: Decide

$$F_o = F_1 \quad \text{if} \quad (u-v)(u+v-1) < 0 \quad .$$

Das Gupta [1964] proved, in his paper, that the above classification procedure based on the Wilcoxon-Statistic is consistent. Kanazawa [1974] proposes the extension of the rule for the multivariate and multisample case, showing its consistency. When the observations are correctly classified, he has shown that his classification statistic is asymptotically distributed according to the chi-square distribution with p (number of variates) degrees of freedom. For details see Kanazawa [1974].

Kinderman [1972] proposed a class of rules based on linear rank statistics as follows: Suppose n observations are available from each

of the three populations π_0, π_1, π_2 . Let $N = 3n$. Define

$$T_{nj} = n^{-1} \sum_{i=1}^{N} E_{Ni} L_{ji} \quad , \quad j = 0,1,2,$$

where E_{Ni} is a sequence of scores and

$$L_{ji} = \begin{cases} 1 & \text{if the } i\text{th ordered observation in the pooled sample is from } \pi_j \\ 0 & \text{otherwise.} \end{cases}$$

Kinderman's rule classifies the observations from π_o into π_1 if and only if

$$2T_{no} - T_{n1} - T_{n2} > 0 \cdot$$

He assumed that the distribution in π_2 differs from that in π_1 by a positive shift in translation. He computed the relative asymptotic efficiency of this rule to the rule obtained by replacing T_{nj} by the corresponding sample mean of the observations from π_j and specialized his results to "Wilcoxon rank-sum" scores and "normal" scores. Govindarajulu and Gupta [1972] consider similar linear rank statistics for the several population case when the sample sizes may be different. For lack of space, the details of these papers are omitted. Interested readers are referred to Kinderman [1972] and Govindarajulu and Gupta [1972].

3.3.4 Best-count Rules:

We discussed in section 3.3.1 the "nearest neighbor" rules which maximize $\tilde{r}$, the counting estimator of the probability of correct classification, over the domain D^* of all discriminants. These are not the only interesting ones related to the counting function $\tilde{r}$. In this section, we shall discuss another of such rules known as "Best-count" rule - a rule which optimizes certain specified criteria in a given class. A systematic study of this concept is due to Glick [1969]. Best-count discriminants generalize sample-based "best" linear or quadratic discriminants.

Consider the set-up as in formulation (ii) of section 2.1. Let $D \subset D^*$, the collection of all discriminants, be arbitrary but a completely specified collection of discriminants D. Then

$$(3.3.12) \qquad r^{D} = \sup_{D \epsilon D} r(D) \quad ,$$

is called the restricted-optimum probability of correct classification.

Definition 3.6: A classification rule $D \epsilon D$ is said to be D-optimal (or restricted optimal for the collection D) if

$$(3.3.13) \qquad r(D) = r^{D} \quad .$$

(In general, there need not exist such a restricted optimum rule.)

Remark 3.2: (i) $r^{D} = \sup_{D \epsilon D} r(D) \leq \sup_{D \epsilon D^*} r(D) = r^*$.

(ii) If, among the classification rules which are optimal in the unrestricted sense, there exists one which is a member of $\mathcal{D}$, then

$$r^* = r(D^*) = r^{\mathcal{D}} \quad .$$

A sample-based rule $\tilde{D} \in \mathcal{D}$ is called a minimum-misclassification discriminant or best-count discriminant if it maximizes $\tilde{r}$ (defined by (3.3.1)), over all $D \in \mathcal{D}$, i.e. a best-count discriminant $\tilde{D} \in \mathcal{D}$ satisfies

$$\tilde{r}(\tilde{D}) = \sup_{D \in \mathcal{D}} \tilde{r}(D) = \tilde{r}^{\mathcal{D}} \quad , \tag{3.3.14}$$

and $\tilde{D}$ is called a best-count rule for the collection $\mathcal{D}$.

Since empirical distributions are simple functions, there necessarily exists a sample-based rule (not usually unique) which maximizes the function $\tilde{r}$ over all the rules $D \in \mathcal{D} \subset \mathcal{D}^*$. It can be noted from the above definition that the nearest neighbor rule $\ddot{D}$, discussed in section 3.3.1, is a best-count discriminant for the collection $\mathcal{D}^*$ of all discriminants. It was seen in section 3.3.1 that for any discriminant D , $\tilde{r}(D)$ is an unbiased estimate of $r(D)$. Using this unbiasedness for a fixed D , Glick [1975] has proved that $E(\tilde{r}^{\mathcal{D}}) = E(\tilde{r}(\tilde{D})) \geq r^{\mathcal{D}} \geq r(\tilde{D})$. He has also proved:

(i) counting function $\tilde{r}$ converges to actual probability of correct classification uniformly over $D \in \mathcal{D}$, i.e.

$$\sup_{D \in \mathcal{D}} |\tilde{r}(D) - r(D)| \xrightarrow{a.s.} 0 \quad \text{as} \quad n \to \infty \ ,$$

provided F_i's are absolutely continuous with respect to the Lebesgue measure.

(ii) The best-count discriminant (or the sequence of such discriminants as sample size $n \to \infty$) is Bayes risk strongly consistent.

He further extends these results to prove that $\tilde{r}(\tilde{D}) \to r^*$, the unrestricted optimum probability, in case of the classification into normal densities with estimated mean vectors and common covariance matrix. (For proofs of these assertions on best-count discriminants see section 4.4 of Chapter IV.)

As a final remark on these best-count discriminants it should be mentioned that the construction of the Fisher-Anderson linear discriminant was explicit in its definition, which is not the case with the best-count discriminants' definition. For arbitrary rule collection, even with $k = 2$ there seems to be no general method for constructing best-count rules (other than by exhaustive trial and error). Glick [1975] remarks that "no general construction of a best-count linear discriminant is yet known when the sample observations from the two populations can not be separated by a hyperplane".

3.3.5 Rules Based on Tolerance Regions.

The idea of using tolerance regions for the classification problem was first suggested by Anderson [1966]. For the univariate case, he considers some variations of NN rules, and in the multivariate case, vector observations may be "ranked" (using them to define blocks) and then a univariate method can be applied. Another method suggested by

Anderson [1966] is: Use the pooled training sample to construct "blocks". An observation is classified into π_i if the blocks to which X belongs is defined by the majority of observations from π_i . Example, for the two population case, construct two sets of blocks separately based on the observations from π_1 and π_2 . Let B_1 and B_2 be the blocks in the two sets which contain X . Consider the number of observations from π_2 in B_1 and the number of observations from π_1 in B_2 , and classify X according to the larger number. The notion of tolerance region is quite important because the expected probability in the region is equal to the number of samples (=k) divided by k+1 . Different methods have been suggested for the construction of tolerance regions. For some details see Patrick [1972].

Quesenberry and Gessaman [1968] also suggested the use of tolerance regions for the k - population nonparametric classification problem with 2^k-1 decisions (instead of k decisions) by introducing the idea of reserve judgment. For details interested readers are referred to their paper.

Remark 3.3: When a statement regarding the probability of a certain statistical decision rule remains valid for every member in a given family of distributions, it is termed as a "Distribution-free" rule with respect to that family. However, in contrast to the problems of hypothesis testing or estimation, nonparametric classification techniques are not really distribution-free. This is because, regardless of the name (parametric, distribution-free or nonparametric), the resulting discriminant function is defined by a set of parameters which must be determined

from the existing prior information. Consequently, we could say that all techniques are somewhat parametric in nature (Andrews [1972], pp. 104).

CHAPTER IV

Mathematical Proofs

In Chapters II and III, we studied various major classification rules (parametric and nonparametric), discussed in the literature. The study also included the sample-based classification rules, the estimates of probability of correct classifications, and mathematical assertions on bias, consistency and asymptotic optimality of these rules. In this chapter, we give mathematical proofs of some of these assertions. Let us recapitulate the different notations that have been used:

(i) $r(D)$ - the actual probability of correct classification for any arbitrary classification rule $D \in \mathcal{D}^*$, the collection of all classification rules, (defined by (2.2.12)).

(ii) $r^* = \sup_{D \in \mathcal{D}^*} r(D)$, the optimal probability of correct classification (defined by (2.2.13)).

(iii) $\hat{r}$ - the density plug-in estimate of the optimum probability of correct classification, r^* (see (2.3.13)).

(iv) for an arbitrary but fixed subcollection $\mathcal{D}$ of $\mathcal{D}^*$, $r^{\mathcal{D}} = \sup_{D \in \mathcal{D}} r(D)$, defines the restricted optimum probability (see (3.3.12)).

(v) $\tilde{r}(D)$ - the counting estimate of the probability of correct classification (see (3.3.1)).

§4.1 Asymptotic Optimality of Density Plug-In Estimators $\hat{r}$.

Theorem 4.1 (Bias): If the estimates $\widehat{q_i f_i}$, $1 \leq i \leq k$ are pointwise unbiased, or more generally if they satisfy:

$$(4.1.1) \qquad E(\widehat{q_i f_i}(x)) \geq q_i f_i(x) \quad , \quad \text{for} \quad 1 \leq i \leq k$$

and for almost all $x \in \mathcal{X}$, then

$$(4.1.2) \qquad E(\hat{r}(\hat{D})) \geq r^* \geq r(\hat{D}) \quad .$$

Proof: The second inequality follows since by definition

$$r^* = \sup_{D \in \mathcal{D}^*} r(D)$$

$$\geq r(\hat{D}) \quad .$$

Further, using $\hat{g}_j(x) = \widehat{q_j f_j}(x)$, $1 \leq j \leq k$, the convexity of $\max\limits_{1 \leq j \leq k} (\cdot)$ and the assumption (4.1.1),

$$E(\hat{\gamma}(x)) = E(\max_{1 \leq j \leq k} \hat{g}_j(x))$$

$$\geq \max_{1 \leq j \leq k} E(\hat{g}_j(x))$$

$$\geq \max_{1 \leq j \leq k} (q_j f_j(x))$$

$$= \gamma(x) \quad .$$

Invoking Fubini's iterated integrals theorem,

$$\begin{aligned} E(\hat{r}(\hat{D})) &= E(\int_{\mathcal{X}} \hat{\gamma}(x)) \qquad \text{(see (2.3.13))} \\ &= \int_{\mathcal{X}} E(\hat{\gamma}(x)) \\ &\geq \int_{\mathcal{X}} \gamma(x) = r^* \ . \end{aligned}$$

(Integration with respect to μ , σ - finite measure, is abbreviated here and often hereafter.)

q.e.d.

Remark 4.1: The usual parametric estimates of multivariate normal densities do not satisfy the conditions of Theorem 4.1. The following is an example (Glick [1972]) satisfying the conditions of Theorem 4.1.

Example 4.1: Consider the counting measure μ on a discrete sample space $\mathcal{X} = \{x_1, x_2, \ldots\}$, (if $\mathcal{X}$ is finite then the distributions are multinomial). Let n_{ik} be the number of individuals from π_i and having $X = x_k$, then n_{ik} is a binomial random variable with expectation $n\ q_i\ f_i(x_k)$. The usual nonparametric density estimates of f_i , $1 \leq i \leq k$, are given by

$$\hat{f}_i(x_k) = \frac{n_{ik}}{n_i} \quad , \qquad x_k \in \mathcal{X} \ .$$

Hence

$$\begin{aligned} \hat{q}_i\ \hat{f}_i(x_k) &= \frac{n_i}{n} \cdot \frac{n_{ik}}{n_i} \qquad \text{(see (2.3.1))} \\ &= \frac{n_{ik}}{n} \end{aligned}$$

and (4.1.1) holds.

The following theorem gives one of the valuable features of the plug-in estimator, $\hat{r}$:

Theorem 4.2 (uniform consistency): If the density estimators $\hat{f}_i$, $1 \leq i \leq k$, are themselves probability densities with respect to a σ - finite measure μ , which converge pointwise with probability one, i.e. if

$$(4.1.3) \quad \hat{f}_i\,(x) \xrightarrow[p]{a.s} f_i(x) \quad \text{and} \quad \int_{\mathcal{X}} \hat{f}_i(x)\, d\mu(x) \xrightarrow{a.s} 1 \ .$$

then

$$(4.1.4) \qquad \sup_{D \in \mathcal{D}^*} |\hat{r}(D) - r(D)| \xrightarrow[p]{a.s.} 0 \ .$$

Proof: Let $D \in \mathcal{D}^*$ be any classification procedure.

$$\begin{aligned} |\hat{r}(D) - r(D)| &= \Big| \sum_{i=1}^{k} \int_{D_i} \hat{g}_i - \sum_{i=1}^{k} \int_{D_i} g_i \Big| \\ &\leq \sum_{i=1}^{k} \int_{D_i} |\hat{g}_i - g_i| \\ &\leq \sum_{i=1}^{k} \int_{\mathcal{X}} |\hat{g}_i - g_i| \ . \end{aligned}$$

This last bound does not depend on the rule D , and hence it also bounds $\sup_{D \in \mathcal{D}^*} |\hat{r}(D) - r(D)|$. Consequently,

$$(4.1.5) \qquad \sup_{D \in \mathcal{D}^*} |\hat{r}(D) - r(D)| \leq \sum_{i=1}^{k} \int_{\mathcal{X}} |\hat{g}_i - g_i| \ .$$

It therefore suffices to show that the integrals

$$\int_{\mathcal{X}} |\hat{g}_i - g_i| \xrightarrow[p]{a.s.} 0 \quad , \quad 1 \leq i \leq k \quad .$$

By (2.3.2) $\hat{q}_i \xrightarrow{a.s.} q_i$, for $1 \leq i \leq k$ and by hypothesis $\hat{f}_i \xrightarrow[p]{a.s.} f_i$ for $1 \leq i \leq k$. These imply the pointwise convergences

$$\hat{q}_i \; \hat{f}_i(x) \xrightarrow[p]{a.s.} q_i \; f_i(x) \qquad \text{(proof trivial)}$$

and

$$\hat{f}(x) \xrightarrow[p]{a.s.} f(x) \qquad \text{(proof trivial)}$$

where $\hat{f}(x) = \sum_{i=1}^{k} \hat{q}_i \; \hat{f}_i(x)$, estimates the mixed density $f(x) = \sum_{i=1}^{k} q_i \; f_i(x)$. Since $0 \leq \hat{g}_i(x) \leq \sum_{i=1}^{k} \hat{q}_i \; \hat{f}_i(x) = \hat{f}(x)$ and $\int_{\mathcal{X}} \hat{f}(x) \, d\mu(x) \xrightarrow[p]{a.s.} 1 = \int_{\mathcal{X}} f(x) \, d\mu(x)$, the desired convergence $\int_{\mathcal{X}} |\hat{g}_i - g_i| \xrightarrow[p]{a.s.} 0$, follows from Lebesgue dominated convergence theorem.

If further, $\int_{\mathcal{X}} \hat{f}(x) \leq 1$ then

$$\begin{aligned} \int_{\mathcal{X}} |\hat{g}_i - g_i| &\leq \int_{\mathcal{X}} \hat{f}(x) + \int_{\mathcal{X}} f(x) \\ &\leq 2 \quad . \end{aligned}$$

And $\sup_{D \in \mathcal{D}^*} |\hat{r}(D) - r(D)| \leq 2k$ (from (4.1.5)).

For an a.s uniformly bounded sequence of random variables, convergence in probability implies convergence in quadratic mean.

q.e.d.

<u>Example 4.2</u>: The following is an example to show that the condition (4.1.3) is vital to Theorem 4.2.

Suppose $\mathcal{X} = (0,1)$ and μ is the Lebesgue measure. Let $q_1 = q_2 = \frac{1}{2}$, $f_1 = f_2 = f = 2\gamma$,

$$\hat{f}(x) = \begin{cases} f(x) & \text{if } x \geq \frac{1}{n} \\ f(x) + n & \text{if } x < \frac{1}{n} \end{cases}$$

and $\hat{\gamma} = \frac{1}{2}\hat{f}$. Then $\hat{\gamma}(x) \xrightarrow{a.s} \gamma(x)$ and $\hat{f}(x) \xrightarrow{a.s} f(x)$ at all $x \in \mathcal{X}$.

Using (2.2.14) and (2.3.13), we have $r^* = \frac{1}{2}$ but $\hat{r} = \frac{1}{2} + \frac{1}{2}\, n\, (\frac{1}{n}) = 1$, for $n = 1,2,3,\ldots$.

Remark 4.2: Theorem 4.2 states general conditions under which $\hat{r}$ is a consistent estimator of r^* . Theorems 4.1 and 4.2 together suggest that $\hat{r}$ is more appropriate as an estimate of r^* , than as an estimate of $r(\hat{D})$. Glick [1973] obtains similar results for sample-based multinomial classification.

Theorem 4.3: Let μ be the Lebesgue measure and if $\hat{q}_i\, \hat{f}_i(x) = \frac{n_i}{n}\, \hat{f}_i(x)$, where each $\hat{f}_i$, $1 \leq i \leq k$, is a Loftsgaarden and Quesenberry density estimate defined by (3.2.5), then the corresponding plug-in estimate $\hat{r}$ satisfies

$$\hat{r} \xrightarrow{p} r^* \quad . \tag{4.1.6}$$

Proof: Theorem 3.1 of Loftsgaarden and Quesenberry [1965] asserts that, for $1 \leq i \leq k$,

$$\hat{f}_i(x) \xrightarrow{p} f_i(x) \qquad \text{at each} \quad x \in \mathcal{X} .$$

Consequently, the assertion of the theorem follows from Theorem 4.2.

q.e.d.

The following is a consistency theorem for general parametric densities considered in section 3.2.

Theorem 4.4: If each $f_i\ (x;\lambda)$ is a continuous function of the unknown parameter λ, for $1 \leq i \leq k$, and for all $x \in \mathcal{X}$, and if

(4.1.7) $$\hat{\lambda} \xrightarrow[p]{a.s} \lambda^* \quad \text{(true value of } \lambda\text{), then}$$

(4.1.8) $$\hat{r} \xrightarrow[p]{a.s} r^* \quad \text{and} \quad \hat{r} \xrightarrow{q.m} r^* \quad .$$

Proof: Continuity of $f_i(x;\lambda)$ and (4.1.7) implies, for $1 \leq i \leq k$,

$$\hat{f}_i(x) = f_i(x;\hat{\lambda}) \xrightarrow[p]{a.s} f_i(x;\lambda^*) = f_i(x)$$

and

$$\int_{\mathcal{X}} \hat{f}_i(x)\, d\mu(x) = \int_{\mathcal{X}} f_i(x;\hat{\lambda})\, d\mu = 1 \quad \text{identically.}$$

Thus, the conclusions follow from Theorem 4.2, since the Lebesgue measure is a σ - finite measure.

q.e.d.

Corollary 4.1: Suppose $k = 2$, and the distributions F_1 and F_2 are multivariate normal with common covariance matrix Σ. If $\hat{f}_1$ and $\hat{f}_2$ are the appropriate multivariate normal density estimators, then $\hat{r}$ satisfies (4.1.8).

Proof: The strong consistency of the parameters follows from the strong law of large numbers, and the conclusions follow from a simple and direct application of Theorem 4.4.

q.e.d.

§4.2 Asymptotic Optimality of Sample-Based Classification Rules.

Once a sample-based procedure is defined, one question that arises is, in what sense is the rule asymptotically optimal. Several modes of asymptotic optimality for classification rules have been proposed in the literature. The following mathematical proofs of asymptotic optimality of parametric and nonparametric classification rules have been adapted from Van Ryzin [1966]. We consider the two category classification problem.

Let q and $1-q$ be the prior probabilities associated with the two populations π_1 and π_2 respectively. Then an optimal Bayes rule, D^*, with respect to these prior probabilities, is given by (see section (2.2.1))

$$D_1^* = \{x \in \mathcal{X} : q\, C_{12}\, f_1(x) > (1-q)\, C_{21}\, f_2(x)\}$$

$$D_2^* = \mathcal{X} - D_1^*; \tag{4.2.1}$$

ties to be resolved in some manner, as discussed in Chapters II and III.

If q, f_1 and f_2 are known, the classification problem is solved by (4.2.1). When f_1 and f_2 are unknown, given random samples of size n_i from π_i, we seek estimates $\hat{f}_i$ for f_i $(i = 1,2)$.

Assume that these samples are independent of the observation X to be classified. Let $\{g_k(x,y) \; ; \; k = 1,2,\ldots\}$ be a sequence of real-valued measurable functions defined on $\mathcal{X} \times \mathcal{X}$ such that a.e. μ

$$(4.2.2) \qquad \int g_k(x,y) \, f_i(y) \, d\mu(y) < \infty \qquad \text{for} \quad i = 1,2 \; ; \; k = 1,2,3,\ldots \; .$$

Then form the estimates

$$(4.2.3) \qquad \hat{f}_i(x) = \frac{1}{n_i} \sum_{k=1}^{n_i} g_{n_i}(x, X_k^{(i)}) \quad , \quad i = 1,2 \; .$$

Assuming these estimates are good in some sense, a reasonable procedure to use in place of (4.2.1) is the plug-in rule, $\hat{D}$, given by

$$\hat{D}_1 = \{x \in \mathcal{X} : q \, C_{12} \, \hat{f}_1(x) > (1-q) \, C_{21} \, \hat{f}_2(x)\}$$

$$(4.2.4) \qquad \hat{D}_2 = \mathcal{X} - \hat{D}_1 \; .$$

<u>Lemma 4.1</u>: The difference in the Bayes risks, $R(\hat{D}) - \rho(D^*)$ satisfies the following inequality:

$$(4.2.5) \qquad 0 \leq R(\hat{D}) - \rho(D^*) \leq C_{12} \, q \int |\hat{f}_1(x) - f_1(x)| \, d\mu(x)$$

$$+ \, C_{21}(1-q) \int |\hat{f}_2(x) - f_2(x)| \, d\mu(x) \quad .$$

<u>Proof</u>: The first inequality follows by the optimality of the Bayes rule, D^* . And the second inequality follows from the expressions for $R(\hat{D})$ and $\rho(D^*)$ given by,

$$\rho(D^*) = \int_{D_2^*} q \, C_{12} \, f_1(x) + \int_{D_1^*} (1-q) \, C_{21} \, f_2(x) \quad ,$$

$$R(\hat{D}) = \int_{\hat{D}_2} q\, C_{12}\, \hat{f}_1(x) + \int_{\hat{D}_1} (1-q)\, C_{21}\, \hat{f}_2(x) \quad ,$$

where D_i^* $(i=1,2)$ and $\hat{D}_i$ $(i=1,2)$ are given by (4.2.1) and (4.2.4).

q.e.d.

Remark 4.3: From Markov's inequality (Loéve [1963] pp. 158), the inequality (4.2.5) and Fubini's theorem we have

$$(4.2.6) \qquad P[R(\hat{D})-\rho(D^*) \geq \epsilon] \leq \epsilon^{-1}\{C_{12}\, q \int E\, |\hat{f}_1(x)-f_1(x)|\, d\mu(x) + (1-q)\, C_{21} \int E\, |\hat{f}_2(x)-f_2(x)|\, d\mu(x)\} \quad .$$

Consequently, it follows that examining Bayes risk consistency (definition 3.1) of rules amounts to studying the asymptotic behaviour of $\int E\, |\hat{f}_i(x)-f_i(x)|\, d\mu(x)$ as $n_i \to \infty$, $i = 1,2$.

In the following theorem, let

$$(4.2.7) \qquad f_i(x) = \sum_{j=1}^{s} \alpha_{ij}\, \psi_j(x) \quad , \quad i = 1,2$$

and for some finite s , where $\psi_j(x)$ are μ - integrable orthonormal functions in $L_2(\mu)$.

Under (4.2.7) we are assuming a parametric form for $f_i(x)$, $(i=1,2)$, but s is assumed to be so large that estimation of α_{ij}'s becomes impractical. Aizerman, Braverman and Rozonoer [1964] use the estimates $\hat{f}_i(x)$, given by (4.2.3) where

$$(4.2.8) \qquad g_k(x,y) = g(x,y) = \sum_{j=1}^{s} \psi_j(x)\, \psi_j(y) \quad , \quad k = 1,2,3,\ldots \quad .$$

These estimates are unbiased for:

$$(4.2.9) \qquad E(\hat{f}_i(x)) = \int g(x,y)\ f_i(y)\ d\mu(y)$$

$$= \sum_{j=1}^{s} \psi_j(x) \int \psi_j(y)\ f_i(y)\ d\mu(y)$$

$$= \sum_{j=1}^{s} \psi_j(x)\ \alpha_{ij} = f_i(x)$$

(using orthonormality of ψ_j's and (4.2.7)).

Theorem 4.5: Under (4.2.7), let $\hat{D}$ be defined by (4.2.3), (4.2.4) and (4.2.8). Then $\hat{D}$ is BRC with D^*.

Proof: Since $\int |g(x,y)|\ f_i(y)\ d\mu(y) < \infty$, by (4.2.9), the strong law of large numbers and L_1 - convergence theorem (Loéve [1963] p. 163), we have

$$E|\hat{f}_i(x)-f_i(x)| \to 0 \quad \text{as} \quad n_i \to \infty .$$

Further,

$$E|\hat{f}_i(x)-f_i(x)| \leq E\ |\hat{f}_i(x)| + f_i(x)$$

$$\leq \sum_{j=1}^{s} |\psi_j(x)| \int |\psi_j(y)|\ |f_i(y)|\ d\mu(y) + f_i(x) \quad ,$$

and the right hand side quantity is μ - integrable. Hence by Lebesgue dominated convergence theorem,

$$\int E\ |\hat{f}_i(x)-f_i(x)| \to 0 \quad \text{as} \quad n_i \to \infty$$

and the conclusions follow from (4.2.6) and remark 4.3.

q.e.d.

We shall state the following theorem (without proof) concerning the asymptotic optimality of nonparametric classification rules, as proved by Van Ryzin [1965, 1966]. Let $\mathcal{X}$ be the Euclidean r - space R^r and μ be r - dimensional Lebesgue measure. We define $\hat{f}_i(x)$ by (4.2.3), by choosing

$$g_k(x,y) = \frac{1}{h_k^r} k\left(\frac{x-y}{h_k}\right) \tag{4.2.10}$$

where $\{h_k\}$ is a sequence of positive numbers satisfying

$$h_k \downarrow 0 \quad \text{as} \quad k \uparrow \infty \tag{4.2.11}$$

and $k(y) = k(y_1, y_2, \ldots, y_r)$ is a bounded Borel function on Euclidean r - space with

$$\int k(y)\, dy = 1 \quad , \quad k(y) \geq 0 \tag{4.2.12}$$

$$||y||^r k(y) \to 0 \quad \text{as} \quad ||y|| \to \infty \, , \; ||y||^2 = \sum_{j=1}^{r} y_j{}^2 \; . \tag{4.2.13}$$

(For a detailed discussion on this density estimation method, see section 3.2.1 and Parzen [1962].)

<u>Theorem 4.6</u>: Let $f_i(x)$ be continuous a.e. with respect to Lebesgue measure μ . Then the rule $\hat{D}$ defined by (4.2.3), (4.2.4) and (4.2.10) is <u>BRC</u> with D^* .

<u>Proof</u>: See Van Ryzin [1965, 1966].

q.e.d.

Following Van Ryzin's notion of Bayes risk consistency (definition 3.1), Glick [1972] proved the asymptotic optimality of the density plug-in rule $\hat{D}$.

Theorem 4.7: Subject to the conditions of Theorem 4.2, the density plug-in rule $\hat{D}$ is Bayes risk consistent (or strongly consistent); and $\hat{r}(\hat{D})$ is a consistent (or strongly consistent) estimator of the optimum probability r^* , i.e.,

$$r(\hat{D}) \xrightarrow[a.s]{p} r^* \quad \text{and}$$

$$\hat{r}(\hat{D}) \xrightarrow[a.s]{p} r^* \quad . \tag{4.2.14}$$

Proof: Theorem 4.2 immediately implies,

$$\hat{r}(\hat{D}) - r(\hat{D}) \to 0 \quad .$$

Moreover,

$$|\hat{r}(\hat{D})-r^*| = |\sup_{D\epsilon\mathcal{D}^*} \hat{r}(D) - \sup_{D\epsilon\mathcal{D}^*} r(D)|$$

$$\leq \sup_{D\epsilon\mathcal{D}^*} |\hat{r}(D)-r(D)| \quad ,$$

so theorem 4.2 implies

$$\hat{r}(\hat{D}) \to r^* \quad .$$

Further,

$$|r(\hat{D})-r^*| \leq |r(\hat{D})-\hat{r}(\hat{D})| + |\hat{r}(\hat{D})-r^*|$$

$$\to 0 \qquad \text{(by first two convergences).}$$

Thus,

$$r(\hat{D}) \to r^* \quad .$$

q.e.d.

Example 4.3: In the previously considered example 4.1,

$$\hat{q}_i \ \hat{f}_i(x_k) = \frac{n_{ik}}{n} \xrightarrow{a.s} q_i \ f_i(x_k)$$

(by strong law of large numbers). Now,

$$\int_{\mathcal{X}} \hat{f}(x) \ d\mu(x) = \sum_k \hat{f}(x_k)$$

$$= \sum_k \sum_i \hat{q}_i \ \hat{f}_i(x_k)$$

$$= \frac{1}{n} \sum_k \sum_i n_{ik} = 1 \quad .$$

Hence theorems 4.2 and 4.7 apply, with convergence almost surely and in quadratic mean. (Indeed, for $k = 2$ distributions on a finite sample space, Glick [1973] has proved that $P[r(\hat{D}) = r^*] \to 1$, with exponential convergence.)

§4.3 Consistency of Minimum Distance Nonparametric Classification Rule.

In section 3.3.2, we discussed the minimum distance nonparametric classification rule, as proposed by Das Gupta [1964]. We give here the mathematical proofs of the assertions made in that section.

Lemma 4.2: For $k = 2$, the following relation holds: For $i = 1,2$

$$r_{ii}(d) \geq f_d(n_1,\tfrac{\beta}{4},F_1)\; f_d(n_2,\tfrac{\beta}{4},F_2)\; f_d(n_o,\tfrac{\beta}{4},F_o = F_i) \tag{4.3.1}$$

$(i = 1,2)$ where $d(F_1,F_2) \geq \beta > 0$ and $r_{ii}(d)$ and f_d are defined by (3.3.7) and (3.3.8) respectively.

<u>Proof</u>: We shall prove for $i = 1$. The proof is analogous for $i = 2$.

By triangle inequality,

$$d(\hat{F}_0,\hat{F}_1) \leq d(\hat{F}_0,F_1) + d(\hat{F}_1,F_1) \quad , \quad \text{and} \tag{4.3.2}$$

$$d(\hat{F}_0,\hat{F}_2) \geq d(\hat{F}_0,F_2) - d(\hat{F}_2,F_2)$$

$$\geq d(F_1,F_2) - d(\hat{F}_0,F_1) - d(\hat{F}_2,F_2) \ . \tag{4.3.3}$$

By (3.3.4),

$$d(\hat{F}_0,\hat{F}_2) - d(\hat{F}_0,\hat{F}_1) = d_{02} - d_{01} \quad . \tag{4.3.4}$$

Combining (4.3.2), (4.3.3) and (4.3.4) we obtain

$$d_{02}-d_{01} \geq d(F_1,F_2) - d(\hat{F}_1,F_1) - d(\hat{F}_2,F_2) - 2d(\hat{F}_0,F_1) \quad . \tag{4.3.5}$$

$d(\hat{F}_0,F_1) < \frac{\beta}{4}$, $d(\hat{F}_1,F_1) < \frac{\beta}{4}$, $d(\hat{F}_2,F_2) < \frac{\beta}{4}$ give from (4.3.5)

$$d_{02} - d_{01} \geq 0 \quad .$$

Consequently,

$$r_{ii}(d) = P[d_{02} - d_{01} > 0 \mid F_0 = F_1]$$

$$\geq P[d(\hat{F}_1,F_1) < \tfrac{\beta}{4} \ , \ d(\hat{F}_2,F_2) < \tfrac{\beta}{4} \ , \ d(\hat{F}_0,F_1) < \tfrac{\beta}{4} \ / \ F_0 = F_1]$$

$$= f_d(n_1,\tfrac{\beta}{4},F_1)\ f_d(n_2,\tfrac{\beta}{4},F_2)\ f_d(n_o,\tfrac{\beta}{4},F_0=F_1)$$

which proves (4.3.1) for $i = 1$.

q.e.d.

Lemma 4.3: For any i , $(i = 1,2,\ldots,k)$

(4.3.6) $$1 - r_{ii}(d) = \sum_{\substack{j=1\\ j\neq i}}^{k} [1 - B_{ij}(d)]$$

where

(4.3.7) $$B_{ij}(d) = P[d_{oj} > d_{oi} \mid F_o = F_i] \quad (i\neq j\ ,\ i=1,2,\ldots,k)\ .$$

Proof: Let E_{ij} be the event $d_{oi} > d_{oj}$. Then

$$\begin{aligned} 1 - r_{ii}(d) &= P[\bigcup_{i\neq j} E_{ij} \mid F_o=F_i] \\ &\leq \sum_{\substack{j=1\\ j\neq i}}^{k} P[E_{ij} \mid F_o=F_i] \\ &= \sum_{\substack{j=1\\ j\neq i}}^{k} [1 - B_{ij}(d)]\ . \end{aligned}$$

q.e.d.

A well-known theorem on Kolmogorov-distance (def. 3.5) states that:

Theorem 4.8: The Kolmogorov-distance is uniformly consistent in the class of all univariate distribution functions.

Proof: See Das Gupta [1964].

q.e.d.

Theorem 4.9: If the distance function d is consistent (uniform) then the minimum distance classification rule $\phi^{(d)}$, defined by (3.3.4), is consistent (uniform), i.e.

$$r_{ii}(d) \to 1 \quad \forall\, i \ (i = 1,2,\ldots,k) \quad \text{as} \quad n_i \to \infty$$

where $r_{ii}(d)$ is defined by (3.3.7).

Proof: Let $d(F_i,F_j) = \ell_{ij}$, $\ell_{ij} > 0$. Then, by lemma 4.2,

$$B_{ij}(d) \geq f_d(n_i,\frac{\ell_{ij}}{4},F_i)\ f_d(n_j,\frac{\ell_{ij}}{4},F_j)\ f_d(n_o,\frac{\ell_{ij}}{4},F_o{=}F_i)$$

d consistent => each of $f_d(n_i,\frac{\ell_{ij}}{4},F_i)$, $f_d(n_j,\frac{\ell_{ij}}{4},F_j)$, $f_d(n_o,\frac{\ell_{ij}}{4},F_o{=}F_i)$ approaches 1 , as $n_o,n_i,n_j \to \infty$ (by definition of $B_{ij}(d)$).

Consequently, the conclusions follow from lemma 4.3.

Similar argument holds for uniform consistency.

q.e.d.

Corollary 4.2: The minimum distance classification rule based on Kolmogorov distance (in the univariate case) is uniformly consistent.

Proof: Follows immediately from theorem 4.8 and theorem 4.9.

q.e.d.

§4.4 Certain Results on Best-Count Discriminants.

There is a direct parallel to the bias theorem 4.1 for best-count discriminants.

Theorem 4.10 (Bias): For any subcollection $\mathcal{D}$ of $\mathcal{D}^*$ and any sample-

based best-count discriminant $\tilde{D} \in \mathcal{D}$, $\tilde{r}(\tilde{D})$ has expected value greater than or equal to the restricted optimum probability, which in turn is greater than or equal to $r(\tilde{D})$, i.e.

$$E(\tilde{r}(\tilde{D})) \geq r^{\mathcal{D}} \geq r(\tilde{D}) \quad . \tag{4.4.1}$$

Proof: Similar to that of theorem 4.1.

q.e.d.

Theorem 4.11 (uniform convergence): As sample size $n \to \infty$, the counting function $\tilde{r}$ converges to the actual probability of correct classification $r(D)$, uniformly over all discriminants D in the subcollection $\mathcal{D}$, i.e.

$$\sup_{D\in\mathcal{D}} |\tilde{r}(D)-r(D)| \xrightarrow[q.m]{a.s} 0 \tag{4.4.2}$$

provided that $F_1,F_2,\ldots,F_k$ are absolutely continuous with respect to the Lebesgue measure μ .

Proof: Using (3.3.1) and (2.2.12) we have

$$\begin{aligned} |\tilde{r}(D)-r(D)| &= \sum_{i=1}^{k} \hat{q}_i \int_{D_i} d\,\hat{F}_i(x) - \sum_{i=1}^{k} q_i \int_{D_i} d\,F_i(x)| \\ &\leq \sum_{i=1}^{k} \{\hat{q}_i \left|\int_{D_i} [d\,\hat{F}_i(x) - d\,F_i(x)]\right| + |\hat{q}_i-q_i|\int_{D_i} d\,F_i(x)\} \\ &\leq \sum_{i=1}^{k} \{\left|\int_{D_i} d\,\hat{F}_i(x) - \int_{D_i} d\,F_i(x)\right| + |\hat{q}_i-q_i|\} \; . \end{aligned}$$

If $\mathcal{H}(u)$ is the collection of all sets which are intersections of at most u half-spaces then either $D_i \in \mathcal{H}(u)$ or $\mathcal{X} - D_i \in \mathcal{H}(u)$ and

$$\left|\int_{D_i} d\,\hat{F}_i(x) - \int_{D_i} d\,F_i(x)\right| \leq \sup_{S\in\mathcal{H}(u)} \left|\int_S d\,\hat{F}_i(x) - \int_S d\,F_i(x)\right| .$$

Since this bound does not depend on the particular discriminant D , it also bounds $\sup_{D\in\mathcal{D}} |\tilde{r}(D)-r(D)|$ and thus,

$$\sup_{D\in\mathcal{D}} |\tilde{r}(D)-r(D)| \leq \sum_{i=1}^{k} \{ \sup_{S\in\mathcal{H}(u)} \left|\int_S d\,\hat{F}_i(x) - \int_S d\,F_i(x)\right|\}$$

(since $|\hat{q}_i-q_i| \xrightarrow{a.s} 0$ by (2.3.2)). So, to conclude the proof, one needs to prove the convergence

$$\sup_S \left|\int_S d\,\hat{F}_i(x) - \int_S d\,F_i(x)\right| \xrightarrow{a.s} 0 .$$

Any asymptotic result of the above form is called a Glivenko-Cantelli convergence of sample measures and has been established by Theorem 2 of Suzuki [1966]. Since for all $D \in \mathcal{D}^*$,

$$|\tilde{r}(D)-r(D)| \leq \tilde{r}(D) + r(D) \leq 2$$

and thus the convergence in quadratic mean follows from almost sure convergence.

q.e.d.

Remark 4.4: Theorem 4.10 asserts that the best-count discriminant $\tilde{D}$ is asymptotically $\mathcal{D}$ - optimal (optimal in the unrestricted sense if $\mathcal{D}$ contains any optimal discriminant).

Corollary 4.3: Subject to the conditions of theorem 4.7 the best-count discriminant $\tilde{D}$ is Bayes risk strongly consistent, i.e.

$$r(\tilde{D}) \to \sup r(D) \qquad \text{with probability one, and}$$

$$\tilde{r}(\tilde{D}) \to \sup r(D) \qquad \text{with probability one.} \tag{4.4.3}$$

Proof: Similar to the proof of theorem 4.7, restricting the classification rules D to the subcollection $\mathcal{D}$ of $\mathcal{D}^*$.

q.e.d.

Here is an example (Glick [1975]) to show that in the case of classification into one of two multivariate normal distributions with common known identity covariance matrix and with estimated mean vectors, even with simple loss structure and equal prior probabilities, the Fisher-Anderson's plug-in linear discriminant is not necessarily a best-count rule for the collection of all linear classifiers, i.e. $\hat{D} \neq \tilde{D}$ in general.

Example 4.4:

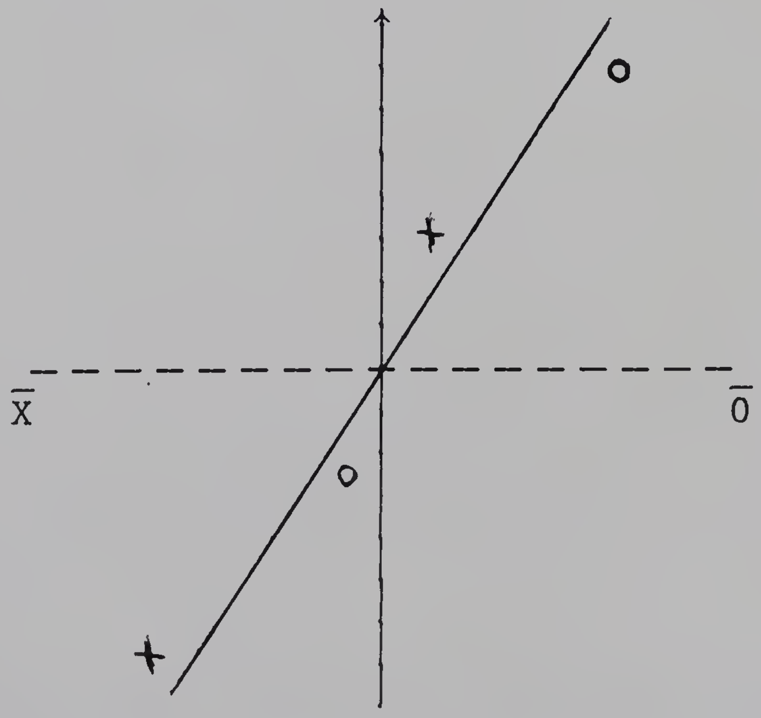

Figure I

Consider a hypothetical sample of $n = 4$ correctly classified bivariate observations from a mixed population Γ . Individuals from π_1 are denoted by X and those from π_2 are denoted by 0 . The solid line (perpendicular bisector of the line segment between sample means) is the Fisher-Anderson Classifier, $(\hat{D})$, and this partition misclassifies one of the three observations from each population. But the diagonal line in Figure I partitions the plane into two disjoint half-spaces and corresponds to a best-count linear discriminant $(\tilde{D})$, which classifies correctly all of the sample points.

CHAPTER V

General Remarks

In the preceding chapters, we discussed various classification procedures - parametric and nonparametric, and some mathematical results on these rules and the associated probabilities of correct classification. In this chapter, we make some general remarks on classification theory, which will be of some use to a statistician.

It was noted that the basic idea in arriving at different classification criteria is the same, namely the rule minimizes the expected loss, or in particular assuming simple loss structure, the probability of misclassification, a natural criterion. After a discriminant or classification procedure has been established, it is of considerable interest to determine whether the discriminant is really useful. The method of studying such a question involves the use of confusion matrix, defined by Massy [1965], which provides a method for summarizing the number of correct and incorrect classifications made by the procedure. One can also investigate the sensitivity of a procedure to deviations from the assumptions under which it was derived. As an example, we mention Lachenbruch [1975]'s Chapter 3, which is concerned with the robustness properties of linear discriminant functions. (For details see Lachenbruch [1975].)

In Chapters II and III, we did not dwell much on classification into one of several populations. There are two reasons for this. Firstly, the essence of the problem is often contained in the two population case, and secondly, the multiple population case may involve more complex

sampling situations. Lachenbruch [1973] has considered two parametric methods for solving such classification problems and studied the relative performance of these two methods using the estimated proportion of correct classification. Kanazawa [1974] developed a nonparametric classification rule based on the Wilcoxon-Statistic for the several population case, proving its consistency.

If the number of p - variates (dimensions) of the problem is too large, the data are subjected to Factor analysis - a technique that attempts to account for the correlation pattern in a set of observable random variables in terms of a minimal number of unobservable random variables called Factors. These fundamental factors and their linear combinations are used to explain the observed data. Evidently, this way some information is lost. Considering the analogy of discriminant analysis with that of regression analysis, it can be said that unlike regression coefficients, discriminant coefficients are not unique, only their ratios are.

In most of the classification procedures, it has been assumed that X , the vector of measurements is readily observable. However, at times it may not be possible to observe every component of X on each unit that is sampled. This gives rise to what is called "incomplete" data. It is worth mentioning that in such cases one may consider a general stochastic process instead of a finite dimensional vector X . The other interesting topics on classification included in the literature are the following:

(i) Sequential Discrimination.

Let $X_1, X_2, \ldots$ be i.i.d. random variables. Observing X's sequentially and knowing that their distribution is one of countably many different probabilities within an arbitrary error level, the general problem of sequential discrimination is: how can we decide which one. Sometimes when the distance between the formulations is fairly small the discriminatory power of the observed variables is insufficient for satisfactory assignment to π_1 or π_2. Several sequential approaches have been proposed to avoid this problem. Suppose that we wish to avoid more than ϵ_1 proportion of errors in π_1 and ϵ_2 in π_2. If it is possible to obtain independent observations on the individual to be classified, then Lachenbruch [1975] suggests the use of sequential probability ratio test to assign to π_1 or π_2.

The variable $U(X)$ of (2.2.8) is normally distributed with mean $\frac{\Delta^2}{2}$ in π_1 and $-\frac{\Delta^2}{2}$ in π_2 and variance Δ^2, where $\Delta^2 = (\mu^{(1)}-\mu^{(2)})' \Sigma^{-1}(\mu^{(1)}-\mu^{(2)})$ is the Mahalanobis generalized squared distance (see section 2.2.3). The assignment rule may be described as follows: A sequential likelihood ratio test of the hypothesis $H_o : X \epsilon \pi_1$ versus $H_1 : X \epsilon \pi_2$ is performed. Observe X_1 and calculate

$$A = \frac{1-\epsilon_2}{\epsilon_1} \quad , \quad B = \frac{1-\epsilon_1}{\epsilon_2}$$

$$\lambda_1 = \frac{f_2(U(X_1);\Delta^2)}{f_1(U(X_1):\Delta^2)} = e^{-U(X_1)} \quad . \tag{5.1}$$

Then, if

$$\lambda_1 \leq B \quad , \quad \text{assign to} \quad \pi_1 \quad \text{and}$$

$$\lambda_1 \geq A \quad , \quad \text{assign to} \quad \pi_2 \ .$$

Otherwise, take a second observation and calculate

$$\lambda_2 = \prod_{i=1}^{2} \frac{f_2(U(X_i);\Delta^2)}{f_1(U(X_i):\Delta^2)}$$

and then compare λ_2 to A and B. This process of taking an observation and calculating λ is continued until λ_i is less than B or greater than A. In general, we have

$$\lambda_i = e^{-\Sigma U(X_i)} = e^{-n\, U(\bar{X})}$$

and consequently, the rule is: Assign to π_1 if after n observations

$$U(\bar{X}) \geq -\frac{1}{n} \ln B$$

to π_2 if

$$U(\bar{X}) \leq -\frac{1}{n} \ln A \ .$$

It is clear that the method described above does not involve prior probabilities q_1 and q_2. This is because we are restricting the individual probability of misclassification. Kendall [1966] suggested a sequential method based on order statistics. The usage of sequential discriminants is not widespread and there is no systematic work on sequential rules. (For more references on this topic see Das Gupta [1973].)

(ii) Logistic Discrimination.

For discriminating between two populations when some or all of observations are qualitative Logistic discrimination was introduced by Cox [1966]. This is found mostly in medical diagnosis based on symptoms and signs and in epidemiology investigating factors related to diseases with low incidences. For more details see Cacoullos [1973, pp. 1-14.]

(iii) Discrimination between Stochastic Processes.

The papers dealing with this problem of discrimination are concerned mainly with finding conditions under which two or more processes (i.e. the induced measures) are equivalent or non-singular. For details see Das Gupta [1973].

(iv) Constrained Discrimination.

In Chapter II, we studied the optimal Bayes rules which minimizes the expected loss or the probabilities of misclassification. However, sometimes, the probabilities of misclassification are so large that the procedure is of little practical use. One alternative is to assign costs to the various types of error which is often difficult or impossible. A second alternative is to decide the probabilities of misclassification within each group that can be tolerated and obtain a rule that satisfies these constraints. These constitute what is called "constrained discrimination".

As is evident, the classification procedures are all strikingly different from one another. Comparisons of different rules in similar

situations should be interesting. In particular, best-count rule and Fisher-Anderson linear discriminant rule might be compared for both normal and non-normal data. Counting estimates of classification probabilities (the R- method of Lachenbruch and Mickey [1968]) have been compared to density plug-in estimates (their D - method) in the case of estimated Fisher-Anderson rule. Lachenbruch and Mickey [1968] conclude that both the estimates are similarly biased for multivariate normal data. The most appealing technique is Anderson's modification of Fisher's linear discriminant, namely, the plug-in linear discriminant, yet he says that it only "seems intuitively reasonable".

Many computer programs are available to perform linear discriminant analyses. The most widely used package is BMD [Dixon, 1974], which has three discriminant analyses' programs, BMD 04M, BMD 05M and BMD 07M. BMD 04M computes a discriminant function for two groups using specified subsets of variables. The output includes group means, covariance matrix, coefficients of the discriminant function and Mahalanobis D^2 . BMD 05M perfoms a multiple-group discriminant analysis for upto five groups. Output includes means, covariance matrix, Mahalanobis' D^2 , coefficients of discriminant functions for each group and a classification matrix. It is assumed that a priori probabilities are the same for each group, which can be a rather serious limitation. BMD 07M performs a stepwise discriminant analysis on upto 80 groups. The variable to enter or to be deleted is selected on the basis of one of three criteria at user's option. Output includes the population means and pooled covariance matrix, classification matrix at specified steps, and posterior probabilities of coming from each population, among others.

This program also has the option of specifying prior probabilities.

It is not difficult to extend the classification framework of this study to cases in which there are k classes and a different finite number L of decision options. Other applications of this generalized framework are suggested by Marshall and Olkin [1968]. Finally, it should be pointed out that the classification problem can be arrived at starting from the framework of Cluster Analysis - whose operational objective is to discover a category structure which fits the observations. In this case little or nothing is known about the category structure, and all that is available is a collection of observations. But, on the other hand, in the case of classification problem, the operational objective is to classify new individuals, i.e. given the category structure, the classification problem amounts to recognising the new individuals as members of one category or another. Cluster Analysis has been employed as a tool in scientific inquiry - a tool of discovery. Biologists give it the name "numerical taxonomy" while the engineers call it "learning without teacher". For a detailed discussion on Cluster Analysis, see Anderberg [1973].

REFERENCES

[1] Aizerman, M.A., Braverman, E.M. and Rozonoer, L.I. The probability problem of pattern recognition learning and the method of potential functions, Automation and Remote Control, 26, (1964), 1175-1190.

[2] Anderberg, M.R. Cluster analysis for applications, Academic Press, N.Y. (1973).

[3] Anderson, T.W.

(a) Classification by multivariate analysis, Psychometrika, 16, (1951), 31-50.

(b) Introduction to multivariate statistical analysis, John Wiley, N.Y. (1958).

(c) Some nonparametric multivariate procedures on statistically equivalent blocks, in Multivariate Analysis, (Ed. P.R. Krishnaiah), Academic Press, N.Y., (1966), 5-27.

[4] Anderson, T.W. and Bahadur, R.R. Classification into multivariate normal distributions with different covariance

[5] Andrews, H.T. Introduction to mathematical techniques in pattern recognition, Wiley, New York, (1972).

[6] Cacoullos, T.

(a) Estimation of multivariate density, Ann. Inst. Stat. Math. (Tokyo), Vol. 18, (1966), 179-189.

(b) Ed. Discriminant analysis and applications, Academic Press, N.Y., (1973).

[7] Cover, T.M. Estimation by the nearest neighbor rule, IEEE trans. Information theory, IT-14, C19681, 50-55.

[8] Cover, T.M. and Hart, F.E. Nearest neighbor pattern recognition, IEEE trans. Information theory, IT-13, (1967), 21-27.

[9] Cox, D.R. Some procedures associated with the logistic qualitative response curve. Research papers in statistics: Festscherift for J. Neyman (F.N. David, Ed.), Wiley, London, (1966) 55-71.

[10] Das Gupta, S.

(a) Nonparametric classification rules, Sankhya, A, 26, (1964) 25-30.

(b) Optimum classification rules for classification into two multivariate normal populations, Ann. Math. Stat., 36, (1965), 1174-1184.

(c) Theories and methods in classification - A Review, in Discriminant analysis and applications, (Ed. T. Cacoullos), Academic Press, N.Y., (1973), 77-137.

[11] Dixon, W.J. BMD, Biomedical computer programs, University of California Press, (1974).

[12] Dunn, O.J. and Varady, P.V. Probability of correct classification in discrimination analysis, Biometrics 22, (1966), 908-924.

[13] Fisher, R.A. The use of multiple measurements in taxonomic problems, Ann. Eugen., 7, (1936), 179-188.

[14] Fix, E. and Hodges, J.L. Nonparametric discrimination consistency properties, U.S. Air Force School of Aviation Medicine, Project no. 21-49-004, report no. 4, Randolph Field, Texas.

[15] Ghurye, S.G. and Olkin, I. Unbiased estimation of some multivariate densities and related functions, Ann. Math. Stat., 40, (1969), 1261-1271.

[16] Glick, N.

(a) Estimating unconditional probability of correct classification, Stanford University, Department of Statistics, Tech. report no. 19, (1969).

(b) Sample-based classification procedures derived from density estimators, Jour. Amer. Stat. Ass., 67, (1972), 116-122.

(c) Sample-based multinomial classification, Biometrics, 29, (1973), 241-256.

(d) Sample-based classification procedures related to empiric distributions, unpublished, (1975).

[17] Govindarajulu, Z. and Gupta, A.K. Certain nonparametric classification rules. Univariate case, Michigan University, Dept. of Stat., Tech. report no. 17, (1972).

[18] Hills, M. Allocation rules and error rates, Jour. Royal Stat. Soc., B, 28, (1966), 1-31.

[19] Kanazawa, M. A nonparametric classification rule for several multivariate populations, Canadian Jour. of Stat., Vol. 2, no. 2, (1974), 145-156.

[20] Kendall, M.G. Discrimination and classification, in multivariate analysis, (Ed. P.R. Krishnaiah), Academic Press, N.Y., (1966), 165-185.

[21] Kshirsagar, A.M. Multivariate analysis, Marcel Dekkar, N.Y. (1972).

[22] Kinderman, A. On some properties of classification: Classifiability, asymptotic relative efficiency, and a complete class theorem, Univ. of Minnesota, Dept. of Statistics, Tech. report no. 178, (1972).

[23] Lachenbruch, P.A.

(a) Some results on the multiple group discriminants, in Discriminant analysis and applications, (Ed. T. Cacoullos), Academic Press, N.Y., (1973), 193-211.

(b) Discriminant analysis, Hafner Press, London, (1975).

[24] Lachenbruch, P.A. and Mickey, D.R. Estimation of error rates in discriminant analysis, Technometrics, 10, (1968), 1-11.

[25] Lehmann, E.L. Testing statistical hypothesis, John Wiley, N.Y. (1959).

[26] Loéve, M. Probability theory, Third Edition, Van Nostrand, (1963).

[27] Loftsgaarden, D.O. and Quesenberry, C.P. A nonparametric estimate of a multivariate density function, Ann. Math. Stat., 36, (1965) 1049-1051.

[28] Mahalanobis, P.C. On the generalized distance in statistics, Proc. Nat. Inst. Sci. India, 2, (1936), 49-55.

[29] Marshall, A.W. and Olkin, I. A general approach to some screening and classification problems, Jour. Royal Stat. Soc. B, 30, (1968), 407-438.

[30] Massy, W.F. On methods: Discriminant analysis of audience characteristics, Jour. of Advertising, res. vol. 5, (1965), 39-48.

[31] Parzen, E. On estimation of probability density and mode, Ann. Math. Stat., 33, (1962), 1065-1076.

[32] Patrick, E.A. Fundamentals of pattern recognition, Prentice Hall, Englewcod Cliffs, N.J., (1972).

[33] Quesenberry, C.P. and Gessaman, M.R. Nonparametric discrimination using tolerance regions, Ann. Math. Stat. 39, (1968), 664-673.

[34] Rao, B.L.S.P. Estimation of unimodal density, Sankhya, A, 31, (1969), 23-36.

[35] Rao, C.R.

(a) Advanced statistical methods in biometric research, Wiley, N.Y., (1952).

(b) Recent advances in discriminatory analysis, Jour. of Indian Soc. of Agri. Stat., 21, (1969), 3-15.

(c) Linear statistical inference and applications, John Wiley, N.Y., Second Edition, (1973).

[36] Srivastava, M.S. Evaluation of misclassification errors, The Canadian Jour. of Stat., Vol. 1, (1973), 35-50.

[37] Suzuki, G. On the Glivanko- Cantelli Theorem, Ann. Inst. Stat. Math (Tokyo), 18, (1966), 29-37.

[38] Van Ryzin, J.

(a) Nonparametric Bayesian decision procedures for (pattern) classification with stochastic learning, Proc. IV Prague Conf. on Inf. Theory, Statistical decision functions and random processes, Academic Press, N.Y., (1965), 479-494.

(b) Bayes Risk consistency of classification procedures using density estimation, Sankhya, A, 28, (1966), 261-270.

(c) On strong consistency of density estimates, Ann. Math. Stat., 40, (1969), 1765-1772.

[39] Wegman, E.J.

(a) Maximum likelihood estimation of a unimodal density function, Ann. Math. Stat. 41, (1970a), 457-471.

(b) Maximum likelihood estimation of a unimodal density, I , Ann. Math. Stat. 41, (1970b), 2169-2174.

(c) Nonparametric probability density estimation - I : summary of available methods, Technometries, 14, (1972), 533-545.

[40] Welch, B.L. Note on discriminant functions, Biometrika, 31, (1939), 218-220.

APPENDIX I

The following data has been taken from the '1975 world population sheet' published by the Population Reference Bureau, Inc. (Washington). The data is based primarily on unpublished United Nations (UN) figures. The data sheet lists all countries with a population larger than 200,000. The variables considered are:

1. Birth rate (= annual number of births per 1,000 population),
2. Death rate (= annual number of deaths per 1,000 population),
3. Life expectancy at birth (years),
4. Per Capita gross national product (US$).

The data for variables 1,2 and 3 come from unpublished materials of the population division of the UN. Birth rates, death rates and life expectancy at birth refer to the average of the 1970-75 period. Per capita gross national product is taken from the International Bank for Reconstruction and Development, 1971 or 1972 data.

The two populations π_1 and π_2 consist of developed and underdeveloped countries (or regions) respectively. The term 'developed' corresponds to low birth and death rates, high life expectancy and reasonably high per capita gross national product. The problem of classification amounts to classifying other countries (namely doubtful) into developed and underdeveloped with respect to these variables. We consider two samples of sizes 30 and 40 respectively from the two populations π_1 and π_2 .

Raw data for the samples from the two populations.

Sample 1 (n_1 = 30).

	(1)	(2)	(3)	(4)
AUSTRALIA	21.0	8.3	72.0	2980.0
AUSTRIA	14.7	12.2	71.0	2410.0
BELGIUM	14.8	11.2	73.0	3210.0
BULGARIA	16.2	9.2	72.0	820.0
CANADA	18.6	7.7	72.0	4440.0
CZECHOSLAVAKIA	17.0	11.2	69.0	2120.0
DENMARK	14.0	10.1	74.0	3670.0
FINLAND	13.2	9.3	70.0	2810.0
FRANCE	17.0	10.6	73.0	3620.0
GERMANY	12.0	12.1	71.0	3390.0
GREECE	15.4	9.4	72.0	1460.0
HUNGARY	15.3	11.5	70.0	1200.0
ICELAND	19.3	7.7	74.0	2800.0
IRELAND	22.1	10.4	72.0	1580.0
ISRAEL	26.5	6.7	71.0	2610.0
ITALY	16.0	9.8	72.0	1960.0
JAPAN	19.2	6.6	73.0	2320.0
LUXEMBOURG	13.5	11.7	71.0	3190.0
NETHERLANDS	16.8	8.7	74.0	2840.0
NEW ZEALAND	22.3	8.3	72.0	2560.0
NORWAY	16.7	10.1	74.0	3340.0
POLAND	16.8	8.6	70.0	1350.0
SINGAPORE	21.2	5.2	70.0	1300.0
SPAIN	19.5	8.3	72.0	1210.0
SWEDEN	14.2	10.5	73.0	4480.0
SWITZERLAND	14.7	10.0	72.0	3940.0

UNITED KINGDOM	16.1	11.7	72.0	2600.0
UNITED STATES	16.2	9.4	71.0	5590.0
USSR	17.8	7.9	70.0	1400.0
YUGOSLAVIA	18.2	9.2	68.0	810.0

Sample 2 (n_2 = 40)

	(1)	(2)	(3)	(4)
ALGERIA	48.7	15.4	53.0	430.0
ANGOLA	47.3	24.5	38.0	390.0
BAHRAIN	49.6	18.7	47.0	640.0
BANGLADESH	49.5	28.1	36.0	70.0
BHUTAN	43.6	20.5	44.0	80.0
BRAZIL	37.1	8.8	61.0	530.0
BURMA	39.5	15.8	50.0	90.0
CHILE	27.9	9.2	63.0	800.0
CHINA	26.9	10.3	62.0	130.0
CONGO	45.1	20.8	44.0	290.0
CUBA	29.1	6.6	70.0	510.0
EGYPT	37.8	14.0	52.0	240.0
ETHOPIA	49.4	25.8	38.0	80.0
FIJI	25.0	4.3	70.0	500.0
HAITI	35.8	16.5	50.0	130.0
INDIA	39.9	15.7	50.0	110.0
INDONESIA	42.9	16.9	48.0	90.0
IRAN	45.3	15.6	51.0	490.0
IRAQ	48.1	14.6	53.0	370.0
JAMAICA	33.2	7.1	70.0	810.0
JORDAN	47.6	14.7	53.0	270.0
KENYA	48.7	16.0	50.0	170.0
LEBANON	39.8	9.9	63.0	700.0
MALAYSIA	38.7	9.9	59.0	430.0

MAURITIUS	24.4	6.8	66.0	300.0
MEXICO	42.0	8.6	63.0	740.0
MONGOLIA	38.8	9.4	61.0	380.0
NEPAL	42.9	20.3	44.0	80.0
NIGERIA	49.3	22.7	41.0	130.0
PAKISTAN	47.4	16.5	50.0	130.0
PERU	41.0	11.9	56.0	520.0
PHILIPPINES	43.8	10.5	58.0	220.0
RHODESIA	47.9	14.4	52.0	340.0
SOUTH AFRICA	42.9	15.5	52.0	850.0
SRI LANKA	28.6	6.4	68.0	110.0
SYRIA	45.4	15.4	54.0	310.0
TANZANIA	50.2	20.1	44.0	120.0
THAILAND	43.4	10.8	58.0	220.0
TURKEY	39.4	12.5	57.0	370.0
UGANDA	45.2	15.9	50.0	150.0

Data of the countries to be classified.

		(1)	(2)	(3)	(4)
1.	Albania	33.4	6.5	69.0	480.0
2.	Argentina	21.8	8.8	68.0	1290.0
3.	Barbados	21.6	8.9	69.0	930.0
4.	Cyprus	22.2	6.8	71.0	1180.0
5.	Hong Kong	19.4	5.5	70.0	980.0
6.	Kuwait	47.1	5.3	67.0	4090.0
7.	Puerto Rico	22.6	6.8	72.0	2050.0
8.	Romania	19.3	10.3	67.0	740.0
9.	Uruguay	20.4	9.3	70.0	760.0
10.	Venezuela	36.1	7.1	65.0	1240.0

(I) Parametric Classification.

Let the populations be normal. The sample means (see table 1), inverse of the estimated covariance matrix, discriminant function coefficients and the Mahalanobis D^2 between the two populations are computed with the help of the computer program BMD 04M (Dixon [1974]).

Table 1

Variable	Mean 1	Mean 2	Difference	Sum
1	17.20995	41.2274	-24.0174	58.43735
2	9.44665	14.43494	-4.98829	23.88159
3	71.66666	53.72499	17.94167	125.39165
4	2600.33325	333.00000	2267.33325	2933.33325

Inverse Matrix of the Estimated Covariance Matrix:

0.00095	0.00076	0.00106	0.0000
0.0076	0.00784	0.00527	0.0000
0.00106	0.00527	0.00420	0.0000
0.0000	0.0000	0.0000	0.0000

Discriminant function coefficients:

-0.00842	0.02610	0.01524	0.00003

Mahalanobis $D^2 = 28.06131$.

(a) Classification using Anderson's rule (2.3.1(i)):

By (2.3.4) classify X into π_1 or π_2 according as,

$$V(X) = X'S^{-1}(\bar{x}^{(1)}-\bar{x}^{(2)}) - \tfrac{1}{2}(\bar{x}^{(1)}+\bar{x}^{(2)})'\ S^{-1}(\bar{x}^{(1)}-\bar{x}^{(2)}) \gtrless 0$$

(assuming equal prior probabilities and equal losses for misclassifications.

1. ALBANIA: $V(X) = -0.0825 < 0$

Therefore Albania is assigned to π_2 .

2. ARGENTINA: $V(X) = 0.0647 > 0$

Hence Argentina belongs to π_1 .

3. BARBADOS: $V(X) = 0.0936 > 0$

Hence Barbados is classified as developed.

4. CYPRUS: $V(X) = 0.0581 > 0$

Hence Cyprus is a developed country.

5. HONG KONG: $V(X) = 0.0074 > 0$

Hence Hong Kong is assigned to population π_1 .

6. KUWAIT: $V(X) = -0.281$

Hence Kuwait is underdeveloped.

7. PUERTO RICO: $V(X) = 0.0787 > 0$

Hence, assigned to population π_1 .

8. ROMANIA: $V(X) = 0.1159$

Thus, Romania belongs to π_1 .

9. URUGUAY: $V(X) = 0.1411$

Hence Uruguay is developed.

10. VENEZUELA: $V(X) = -0.1779$

Hence, assigned to population π_2 .

(b) Classification using Mahalanobis D^2 (2.3.1(ii)):

By (2.3.7) assign to π_1 or π_2 according as,

$$(X-\bar{x}^{(1)})'\ S^{-1}(X-\bar{x}^{(1)}) \lessgtr (X-\bar{x}^{(2)})'\ S^{-1}(X-\bar{x}^{(2)})$$

where the l.h.s. denotes the distance of X from the 1st sample and the r.h.s. denotes the distance of X from the 2nd sample. Let these distances be denoted by D_1 and D_2 respectively.

1. ALBANIA: $D_1 = 0.2657$, $D_2 = 0.0952$.

Thus, Albania belongs to π_2 .

2. ARGENTINA: $D_1 = 0.0645$, $D_2 = 0.1692$

Hence, assigned to population π_1 .

3. BARBADOS: $D_1 = 0.0374$, $D_2 = 0.2245$

Hence, Barbados is developed.

4. CYPRUS: $D_1 = 0.0719$, $D_2 = 0.1882$

Hence, belongs to π_1 .

5. HONG KONG: $D_1 = 0.1868$, $D_2 = 0.2016$

Hence, Hong Kong is developed.

6. KUWAIT: $D_1 = 0.7949$, $D_2 = .2327$.

Thus, Kuwait is underdeveloped.

7. PUERTO RICO: $D_1 = 0.0558$, $D_2 = 0.2132$.

Therefore, is a member of π_1 .

8. ROMANIA: $D_1 = 0.0414$, $D_2 = 0.2731$.

Hence Romania is classified into π_1 .

9. URUGUAY: $D_1 = 0.0121$, $D_2 = 0.337$

Hence, is assigned to π_1 .

10. VENEZUELA: $D_1 = 0.3994$, $D_2 = 0.0436$

Hence, Venezuela is underdeveloped.

(II) Nonparametric Classification by Nearest Neighbor Rule:

First the data was subject to standardization with respect to the mean and standard deviation of each variable (see table 2). Euclidean distance has been considered.

Table 2

Variable	Mean	Standard Deviation
1	30.93	13.36
2	12.3	5.03
3	61.41	11.28
4	1304.71	1374.1

By definition 3.1, an observation $x_n' \in \{X_1,...,X_n\}$ is nearest neighbor to $X = x$ (observation to be classified) if

$$\min_{1 \le i \le n} d(x_i,x) = d(x_n',x) \quad .$$

Since the computations are tedious, we give classifications of only 3 or 4 countries. Other classifications are similar.

(1) ALBANIA: It is nearest neighbor to 'Cuba' which belongs to π_2 (Dist. = 0.335). Hence Albania is classified into π_2 .

(3) BARBADOS: Nearest neighbor to Yugoslavia (Dist. = 0.289) and hence 'Barbados' is developed.

(8) ROMANIA: Nearest neighbor to Yugoslavia (Dist. = 0.255) and hence belongs to π_1 .

(10) VENEZUELA: Nearest neighbor to Jamaica (Dist. = 0.584) and hence is underdeveloped.

Remark: With 'KUWAIT' we get the same minimum distance from Switzerland and Jamaica, so we can arbitrarily assign to π_2 . (Minimum distance = 2.638).

Thus, it is clear, that all the three rules considered give the same classification of the countries to be classified.